# STATISTICAL TEST THEORY
# FOR EDUCATION AND PSYCHOLOGY

Dato N. M. de Gruijter

&

Leo J. Th. van der Kamp

October 2005

© D. N. M. de Gruijter & L. J. Th. Van der Kamp

## CONTENTS

# PREFACE

From the start of the twentieth century when in France Alfred Binet developed the first intelligence test, the importance of the psychological test in western society has grown. The use of tests in research as well as in clinical and other applications mostly has to do with the question to which extent reliable and valid discriminations between persons are possible. Special attention to testing is given within the context of educational applications.

One of the most important topics in test theory is the subject of reliability and validity. We will thoroughly discuss the concept of *reliability*. It will become clear that reliability is a population dependent property: the reliability of a test does not only depend on the quality of the test, but also on the variation within the relevant population.

In classical test theory the test score is a combination of a true score and measurement error. It is possible to define the measurement error in several ways depending on the way one would like to generalize to other testing situations. Generalizability theory, developed from 1963 onwards by Cronbach and his coworkers, effectively deals with this problem. It gives a framework in which the various aspects of test scores can be dealt with. Of much importance to test theory has been the development of item response theory, or IRT for short. In an item response model, or IRT model, the item is the unit of analysis instead of the test. In IRT models the variance of measurement errors is a function of the level or ability of the respondent, an important characteristic which in most classical test theory models is not available in a natural way. IRT has resulted in improvements in test theoretical applications and in new applications as well, for example in *computerized adaptive testing,* CAT for short.

This manuscript has been written for advanced undergraduate and graduate students in psychology, education and other behavioral sciences. The prerequisites are a working knowledge of statistics including the basic concepts of the analysis of variance, regression analysis and some knowledge of estimation theory and methods. Of course, the more background in research methodology and statistical data-analysis the reader has, the more he/she can profit. This text is also meant for researchers in the field of psychological and educational measurement, not typically specialized in mental test theory. It not only pretends a broad overview, but also a critical survey with hopefully knowledgeable comments and criticism on the test theories. An attempt is made to follow recent developments in the field. As aids in instruction, studying and reading, each chapter concludes with exercises the answers of which are given at the end of the book. It is also endeavored to include examples and exhibits where it seemed desirable.

There are some great books on mental test theory around. Gulliksen (1950) and Lord & Novick (1968) should be mentioned first and with great deference. These are the godfathers of classical test theory and they were the ones to codify it. Would generalizability theory have been developed without the work of Lee J. Cronbach? (see e.g. Cronbach, Gleser, Nanda, & Rajaratnam, 1972). As in many fields of science, inventions and developments are not one-man's achievement. So it is with item response theory, and therefore, being aware of doing injustice to other authors, we only mention Rasch (1960), Birnbaum (1968), Lord & Novick (1968) and Lord (1980).

With which other texts does the present text compete? With Roderick P. McDonald's *Test theory: A unified treatment* (1999) in the first place. The latter book gives a thorough introduction into test theory with special emphasis on unidimensional and multidimensional item response theory and on a multivariate approach (factor analysis and structural equation modeling) to test research and development. McDonald also endeavors to unify the several approaches at the end of his book. Our text differs at least in the sense that our aim is to give each of the approaches (classical test theory, generalizability theory and item response theory) its fair deal, and to include the most recent research work. In comparison with McDonald's

*Test theory* we do not pay explicit attention to the multivariate approach to test theory and development; structural equation modeling and multidimensional item response theory are only mentioned in passing. Embretson and Reise's *Item response theory for psychologists* (2000) demonstrates the use of IRT models with ample examples. Our text is not a "how to do it" guide; many such books are available. Examples of recent textbooks in the field of psychological testing are Gregory (2000), Murphy & Davidshofer (2001), and Walsh & Betz (2001). These books are mainly subject-matter oriented; they stress the measurement of ability, interests, and personality, and also emphasize the use of tests in making decisions in education, industry and in clinical testing. Often these books stress the practicalities, while ours pretends to give more insight into the basics and what is behind all that. So our text is more measurement-oriented, highlighting statistical theories of educational and psychological measurements. This is not to say that the present text bears no relevance for the applications of psychological and educational measurements. It not only heavily leans on the various editions of *Educational measurement*, the later editions of the *Standard for psychological and educational testing* (APA, AERA, & NCME) served as guidelines and ample reference is made to them.

Dato N. M. de Gruijter January 2002
Leo J. Th. Van der Kamp

# 1 MEASUREMENT AND SCALING

## 1.1 Introduction

In educational and psychological research and practice the use of measurement procedures or tests is ubiquitous. Such measurement instruments are used for all kinds of psychological and educational assessments. The main types of psychological and educational tests are: intelligence tests, aptitude tests, achievement tests, personality tests, interest inventories, behavioral procedures and neuropsychological tests. The use of such tests is not restricted to psychology and education, but stretches also over other disciplines of the behavioral sciences, and even outside, e.g., in the field of psychiatry. Using tests involves some kind of measurement procedure, and in addition statistical theories for characterizing the results of the measurement procedures, that is, for modeling test scores.

In this chapter we shall first give a broad and generally accepted definition of a test. Then a sketchy introduction will be given in measurement and scaling. Measurement not only pervades daily life, it also is the cornerstone of scientific inquiry. After defining the concept of measurement, scales of measurement and the relation between measurement and statistics will be presented. Some remarks will be made on scales of measurement in relation to the test theory models given later, while the concept of dimensionality of tests will also be discussed.

## 1.2 Definition of a Test

A test is best defined as a standardized procedure for sampling behavior and describing it with categories or scores. This broad definition includes also checklists, rating scales and observation schemes. The essential features of a test are that it is:

- a standardized procedure, i.e. the procedure is administered uniformly over a group of persons.
- a focussed behavioral sample, i.e. the test is focussed at a well-defined behavioral domain. Examples of domains in educational measurement are achievement in arithmetic, or language performance. Psychological tests may also be targeted to constructs or theoretical variables, e.g., depression, extraversion, quality of life, emotionality, and the like, so, at variables that are not directly observable. In other words, such a measurement approach assumes that there exists a psychological attribute to measure. Such a psychological attribute usually is a core element of what is called a nomological network, which maps its relations with other constructs, and also clarifies its relations with observables, i.e. relevant behavior in the empirical world.
- a description in terms of scores or mapping into categories. Using tests implies a form of measurement whereby performances, characteristics, traits are represented in terms of numbers or classifications.

In addition to these features one should also mention that once a test score is obtained, norms or standards of a relevant group of persons are necessary for the interpretation of the score of a given person. Finally, collecting test scores is seldom an aim in itself, the function of testing is ultimately decision making in a narrow as well as in a broad sense. This includes: classification, selection and placement, diagnosis and treatment planning, self-knowledge, program evaluation, research.

## 1.3 Measurement and Scaling

Stevens (1968) defined measurement as "the assignment of numbers to aspects of objects or events according to one or another rule or convention" (Stevens, 1968, p. 850). Other, sometimes broader, sometimes more refined and more sophisticated definitions are around, but for our purpose Stevens' definition suffices. In addition to what is called psychometric measurement considered here, representational measurement has been formulated. More can be found in Judd & McClelland (1998) and the references mentioned by them, or Michell (1999), who provides a critical history of the concept and in McDonald (1999), who discusses measurement and scaling theory in the context of a unified treatment of test theory.

Usually a test consists of a number of items. The simplest item type is the one where only two answers are possible, e.g. *Yes* or *No*, *correct* or *incorrect*.

After a test has been administered to a group of persons, we generally have a score for each person. The simplest example of a test score is the total score on a multiple choice test where 1 point is given for a correct answer to an item and 0 points to an incorrect answer or skipped item. Some persons have higher scores than others and we expect that these differences are relevant.

We speak of a measurement once a score has been computed. The measurement refers to a property or aspect of the person tested. A well-known classification of measurement scales is given by Stevens (1951). These measurement scales are:

1 the nominal scale
   On the nominal scale objects are classified according to a characteristic.
   An example: one can classify persons with respect to sex, hair color, etc.

2 the ordinal scale
   On the ordinal scale objects are ordered according to a certain characteristic.
   An example: the Beaufort scale of wind force.

3 the interval scale
   On the interval scale equal scale differences imply equal differences in the relevant property.
   Example: the Celsius and Fahrenheit scales for temperature are interval scales; a difference of 1° at the freezing point is as large as a difference of 1° at the boiling point of water.

4 the ratio scale
   The ratio scale has a natural origin as well as equal intervals. Length in meters and weight in kilograms are defined on a ratio scale; so is temperature on the Kelvin scale. Ratio scales are relatively rare in psychology because of the difficulty of defining a zero point. How would persons look like with zero intelligence?

Most researchers do not regard the use of the nominal scale as measurement. One should at least be able to make a statement about the amount of the property in question. Many researchers use an even narrower definition of measurement: they restrict themselves to scales that have at least interval properties.

With interval measurements of temperature two scales are in use: the Celsius scale and the Fahrenheit scale. The scales are related to each other through a linear transformation:

$$^{\circ}\mathrm{F} = (9/5)\ ^{\circ}\mathrm{C} + 32.$$

The linear transformation is a permissible transformation: with a linear transformation the interval properties of the scale are maintained. When we have a ratio scale a general linear transformation is not permissible while such a transformation effects a change of the origin (0). With a ratio scale only multiplication with a constant is permitted. For example, one can measure length in centimeters instead of meters. With an ordinal scale all monotonously increasing transformations are permitted.

The scale properties are relevant when one wants to compute measures characterizing distributions and apply statistical tests. When an ordinal scale is used, one generally is not interested in the average score. The median seems more appropriate and useful. On the other hand, statistics seldom is interested in the measurement level of a variable (Anderson, 1961). When a statistical test is used, it is important to know whether the distributional assumptions hold. Even in case the assumptions are not fully met, statistical tests may be used if these tests are robust against violations of the assumptions.

The interpretation of the outcome of a statistical test, however, does depend on the assumption with respect to the measurement level (Lord, 1954). And, as in some cases a nonlinear transformation might reverse the order of two means, we should decide which kind of transformations we are prepared to apply and which kind of transformations we judge as too extreme to be relevant. More on measurement scales and statistics is presented in Exhibit 1.1.

**Exhibit 1.1. On Measurement Scales or "What to do with football numbers"**

How devoted must a researcher be to Stevens' measurement-directed position? Is it permitted to calculate means and standard deviations on scores on an ordinal scale? Lord (1953) relates a story about a professor who retired early because of feelings of guilt for calculating means and standard deviations of test scores. The university gave this professor the concession for selling cloth with numbers for football players, and a vending machine, to assign numbers randomly. The team of freshman football players protested after a while, because the numbers given to them were too low. The professor consulted a statistician. What to do in the dispute with the complaining members of the freshman football team? Are their football numbers sold indeed too low? The daring and realistic statistician without any hesitation whatsoever, turned to compute all kinds of measures including means and standard deviations of football numbers. The professor protested: these football numbers did not even constitute an ordinal scale! The statistician, however, retorted: "The numbers don't know that. Since the numbers don't remember where they come from, they always behave just the same way regardless" (Lord, 1953, p. 751). The statistician concluded that it was highly implausible that the numbers of the team were a random sample. Needless to say, Lord's professor turned out to be convinced and lost his feelings of guilt. He even took up his old position.

Lord's narrative is basic to the so-called measurement-independent position. However "the utmost care must be exercised in interpreting the results of arithmetic operations upon nominal and ordinal numbers; nevertheless, in certain cases such results are capable of being rigorously and usefully interpreted, at least for the purpose of testing a null hypothesis" (Lord, 1954, p. 265).

In practice we may generally assume that the score scales of psychological and educational tests are not interval scales. Nevertheless researchers frequently act as if the score scale is an interval scale. One might say that no harm is done as long as the predictions from this way of interpreting test results are useful. When difference scores are used as an indication of a learning result or an improvement and these scores are related to other variables, certainly the interval property is invoked. In other test theoretical applications, for example in nonlinear equating of tests – here tests differing in difficulty level and other scale aspects are scaled to the same scale – implicitly the interval property is rejected. In item response models scores on different tests are nonlinearly related to each other. With these models scores can be computed on a latent scale and within the context of a particular model the scale has the interval property. The remaining question is whether this interval property is a fundamental property of the characteristic or just a property that is a consequence of the scale representation chosen. The Rasch model for example has two representations of the characteristic measured, one representation on an additive scale (which is a special case of the interval scale) and a representation with a multiplicative model.

In many applications it is assumed that one dimension underlies the responses to the items of the test in question (see Exhibit 1.2). In principle, in intelligence testing e.g., various abilities interplay in the process of responding to the test item. Let us give an example. In order to be able to respond correctly to mathematics items the persons or examinees in the target population must be able to read the test instructions. Reading ability is needed, but it can be ignored because it does not play a role in the differences between persons tested. Some authors, however, argue that responses always are determined by more than one factor. In ability testing factors like speed, accuracy and continuance have a role (Furneaux, 1960).

### Exhibit 1.2. Dimensionality of Tests and Items

Once measurement became common practice in scientific research in the behavioral sciences, the concept of dimensionality, or more specifically the concept of unidimensionality, emerged as a crucial requirement for measurement.

Two early psychometricians, Thurstone and Guttman, already stressed the importance of unidimensionality for constructing good measures, without using the term though:

> "The measurement of any object or entity describes only one attribute of the object measured. This is a universal characteristic of all measurement"(Thurstone, 1931, p. 257).
>
> "We shall call a set of items of *common content* a scale if (and only if) a person with a higher rank than another person is just as high or higher on *every* item than the other person"(Guttman, 1950, p. 62).

After these early voices not much attention has been paid to a systematic treatment of (uni)dimensionality of items and tests. Exceptions are Levy (1973), McDonald (1981, 1999) and Hattie (1985). A test of unidimensionality has been proposed by Stout (1987) and generalized by Nandakumar, Yu, Li & Stout (1998). The notion of unidimensionality should be discerned from, e.g., internal consistency. A multidimensional set of items may yield a high coefficient of internal consistency (coefficient alpha). Most of the time some sort of factor analysis is used, other methods of dimensional analysis are around (see e.g., Jacoby (1991) for an introduction to dimensional analysis other than factor analysis; see also the methodology review on assessing unidimensionality of tests and items by Hattie (1985)). In the context of item response theory unidimensionality is defined as the existence of one latent trait underlying the data.

In classical test theory no explicit assumption is made with respect to the dimensionality of tests. Some tests are useful just because the items are not restricted to a small domain of unidimensional items, but belong to a broader, more articulated domain of interest. In generalizability theory the possibility to generalize to a heterogeneous domain of reactions is explicitly present. In an anxiety questionnaire one might, for example, ask whether anxiety is raised in a number of different situations and it is assumed that for respondents anxiety is partly situational. But if a researcher is interested in growth or change, test dimensionality is an important issue. For, if the test responses are determined by more than one dimension it is not clear which dimension is responsible for a change in the test responses.

Even in case it can be deduced from test results that the test is unidimensional, one should not conclude that one trait or characteristic determines the responses. One should not mistakenly conclude from a consistency in responses that respondents actually possess a particular trait. When we speak here of abilities or (latent) traits, this is meant for the sake of succinctness: the responses can be described as if the respondents possess a certain latent trait.

In the one-dimensional item response models that will be discussed the responses to the different test items are a measure for an underlying latent trait, i.e. the expected score is an increasing function of the underlying trait. In this context the test items as well as the persons are positioned on the underlying trait or dimension. This is also called the scaling or mapping of items and persons on the same underlying dimension.

## Exercises

1.1 Two researchers evaluate the same educational program. Researcher A uses an easy test as a pretest and posttest, researcher B uses a relatively difficult test. Is it likely that their results will differ? If that is the case, in which way are the results expected to differ?

1.2 In a tennis tournament five persons play in all different combinations. A wins all games, B wins from C, D and E, C wins from D and E, and D wins from E. The number of games won is taken as total score. Which property has this score?

# 2 CLASSICAL TEST THEORY

## 2.1 Introduction

It is a trite observation that all human endeavor is replete with error. And human endeavor called science is no exception. We err in our measurements, that is to say: how hard we may try, never will our measurements be perfect. “O heaven! Were man but constant, he were perfect. That one error fills him with faults; makes him run through all the sins. Inconstancy falls off ere it begins” (Shakespeare: The two gentlemen of Verona. Act v.iv.110-114.). Inconsistency is not the only error though.

There are many possible ways to err in measurement. In other words, there are many sources of errors. These sources may vary dependent upon the particular branch of science involved. The question now is to tackle the problem of errors of measurement. The answer to this question appears to be simple: develop a theory of errors, or some would say, set up an error model. Indeed, in psychological and educational measurements, this is an approach that has been followed since more than a century. And the earliest theory around is classical test theory.

In this chapter classical test theory is presented. By defining true score an explicit, abstract formulation of measurement error is given. This will be the theme of the next section. In Section 2.3 further details will be given on the population of subjects or persons, a topic relevant for further developing test theory, more specifically, for deriving reliability estimates. Also the central assumptions of classical test theory will be given. These are relevant for reliability also, and for considering various types of equivalence or comparability of test forms.

## 2.2 True Score and Measurement Error

Suppose that we have obtained a measurement $x_{pi}$ on person $p$ with measurement instrument $i$. Let us assume, for example, that we have read the weight of this person from a particular weighing-machine and registered the outcome. Next, we take a new measurement and we notice a difference with the first. The obtained measurements can be thought of as arising from a probability distribution for measurements $X_p$ with realizations $x_p$.

With measurement in psychology and education we have a similar situation. We obtain a measurement and we expect to find another outcome from the measuring procedure if we would be able to repeat the procedure and replicate the measurement result. However, in psychology and education we frequently are not able to obtain a series of comparable measurement results with the same measurement instrument because the measurements may have their impact on the person from whom measurements are taken. Memory effects prevent independent replications of the measurement procedure. We might, however, administer a second test constructed for measuring the same construct and notice that the person obtains a different score on this test than on the first test. So, here comes in the development of an appropriate theory of errors or error model. The simplest one is the following. The underlying idea is that the observed test score is contaminated by a measurement error. The observed score is considered to be composed of a true score and a measurement error (see also Figure 2.1):

$$x_p = \tau_p + e_p. \tag{2.1}$$

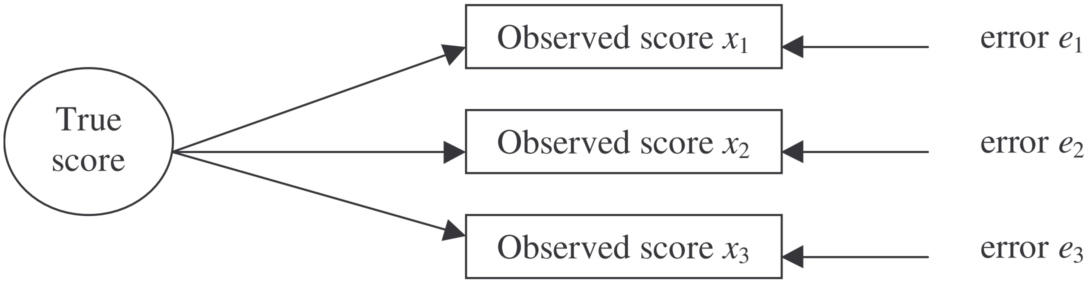


**FIGURE 2.1.** The decomposition of observed scores in classical test theory

If the measurement could be repeated many times under the condition that the different measurements are experimentally independent, then the average of these measurements would give a reasonable approximation to $\tau_p$. In formal terms, true score is defined as the expected value of the variable $X_p$ ($x_p$ from (2.1) is a realization of the random variable $X_p$):

$$\tau_p = \mathrm{E}X_p, \tag{2.2}$$

where E represents the expectation over independent replications.

**Exhibit 2.1. Measurement Error: Systematic and Unsystematic**

Classical test theory assumes unsystematic measurement errors: this type of error is elaborated in Section 2.2. Systematic measurement error may occur when a test consistently measures something other than the test purports to measure. A depression inventory, for example, may not merely tap depression as the intended trait to measure, but also anxiety. In this case a reasonable decomposition of observed scores on the depression inventory would be

$$X = \tau + E_D + E_U,$$

where $X$ is the observed score, $\tau$ is the true score, $E_D$ is the systematic error due to the anxiety component, and $E_U$ the combined effect of unsystematic error.

Clearly, the decomposition of observed score according to classical test theory is only the most rudimentary form of a so-called linear model decomposition. Generalizability theory (Chapter 5) has to say more on the decomposition of observed scores. Also structural equation modeling might be used to unravel the components of observed scores.

The definition of true score as an expected value seems obvious in case the measurements to be taken can be considered exchangeable. In other words, this definition seems obvious if we do not know anything about a particular measurement. But consider the situation in which different measurement instruments are available and we have information on these instruments. Let us give an example. Assume we have some raters as measurement instruments. Assume also that the raters differ in leniency, a fact known to occur. Does the definition of true score as an expected value do justice to this situation? Shouldn't we correct the scores given by a rater with a known constant bias? The answer is that we can correct the scores without rejecting the idea of a true score. For, it is possible to use the score scale of a particular rater and define a true score for this rater. Scores obtained on this scale can be transformed to another scale, comparable to the transformation of degrees Fahrenheit into degrees Celsius. The transformation of scores to scales defined by other measurement instruments will be discussed in the chapter on equating (Chapter 10).

In other situations the characteristics of a particular rater are unknown. Neither is it deemed necessary to have information on this rater, because the next measurement is likely to be taken by another rater. Then the rater effect can be considered part of the measurement error. In Exhibit 2.1 more information on multiple sources of measurement error is given.

The foregoing means that the definition of measurement error, and consequently the definition of true score depends on the situation in which measurements are taken and used. If a particular aspect of the measurement situation has an effect on the measurements and if this aspect can be considered as fixed, one can define true score so as to incorporate this effect. This is the case when one tries to minimize noise in the data to be obtained through the testing procedure by standardization. In other cases one is not able or not prepared to fix an aspect, and the variation due to fluctuations in the measurement context is considered part of the measurement error.

Classical test theory can deal with only one true score and one measurement error. Therefore the test researcher or test user must formulate precisely which aspects belong to the true score and which to measurement error. This choice also restricts the choice of methods to estimate reliability, which is the extent to which obtained score differences reflect true differences. Suppose we want to measure a characteristic which fluctuates from day to day, but which also is relatively stable in the long term. We might be interested in the momentary state, or in the expectation on the long term. If we are interested in measuring the momentary state the value of the test-retest correlation has not much relevance. A systematical framework for the many aspects of measurement errors and true scores has been developed in generalizability theory.

From the definition of true score we can deduce that the measurement error has an expected value equal to zero 0,

$$\mathrm{E}E_p = 0. \tag{2.3}$$

The variance of measurement errors equals:

$$\sigma^2(E_p) = \sigma^2(X_p). \tag{2.4}$$

The square root from the variance in (2.4) is the standard error of measurement for person $p$, the person-specific standard error of measurement.

## 2.3 The Population of Persons

Until this moment we have treated measurements restricted to one person. In practice we usually deal with groups of persons. If a person is tested, the test score is always interpreted within the context of measurements previously obtained from other persons. Test theory is concerned with measurements defined within a population or subpopulation of persons. An intelligence test for example is meant to be used for persons within a given age range, able to understand the test instructions. A population can be large or small.

Selecting a person randomly from the population we have – analogous to (2.1) –

$$X = \mathrm{T} + E, \tag{2.5}$$

where T - the Greek capital tau – designates the true-score random variable.

From the definitions given, the four central assumptions of classical test theory follow:

I The expected measurement error equals 0 (we take the expectation of the person-specific distribution of measurement errors over the population):

$$\mathrm{E}_p E_p = 0 \tag{2.6}$$

II The correlation ρ between measurement error and true score is 0 in the population

$$\rho(\mathrm{T}, E) = 0. \tag{2.7}$$

This follows from the fact that the expected measurement error is equal (equal to 0) for all values $\tau$.

We also assume that two measurements $i$ and $j$ are *experimentally independent*. From this assumption (actually from the weaker assumption of linearly experimental independence) we can deduce III and IV.

III For two measurements $i$ and $j$ holds that the true score on one measurement is uncorrelated with the measurement error on the second measurement:

$$\rho(\mathrm{T}_i, E_j) = 0. \tag{2.8}$$

IV Moreover, the measurement errors of the two measurements are uncorrelated:

$$\rho(E_i, E_j) = 0. \tag{2.9}$$

For the population of persons we also can deduce the equality of the observed population mean and the true-score population mean:

$$\mathrm{E}X = \mathrm{ET} = \mu_X = \mu_\mathrm{T}. \tag{2.10}$$

The result in (2.10) is obvious as well as important. In (2.10) expectations are involved. The observed mean of a small (sub)population certainly is not equal to the true-score mean: the average measurement error may be small, but is unlikely to be exactly equal 0.

The variance of measurement errors can be written as

$$\sigma_E^2 = \mathrm{E}_p \sigma^2(E_p),$$

and the variance of observed scores as

$$\sigma_X^2 = \sigma_\mathrm{T}^2 + \sigma_E^2 + 2\sigma_E \sigma_\mathrm{T} \rho_{TE}.$$

The correlation between true score and error is equal to zero, so we can write the variance of observed scores as:

$$\sigma_X^2 = \sigma_\mathrm{T}^2 + \sigma_E^2. \tag{2.11}$$

The observed-score variance equals the sum of the variance of true scores and the variance of measurement errors.

## Exercises

2.1 A large testing agency administers test *X* to all candidates at the same time in the morning. Other test centers organize sessions at different moments. Give alternative definitions of true score.

2.2 Two intelligence tests are administered close after one another. What kind of problem do you expect?

# 3 CLASSICAL TEST THEORY AND RELIABILITY

## 3.1 Introduction

Classical test theory gives the foundations of the basic true-score model, as discussed in Chapter 2. In the present chapter we shall first go into some properties of the classical true-score model and define the basic concepts of reliability and standard error of measurement (Section 3.2). Then the concept of parallel tests will be discussed; reliability estimation will be considered in the context of parallel tests (Section 3.3). Defining the reliability of measurement instruments is theoretically straightforward, estimating reliability on the other hand requires taking into account explicitly the major sources of error variance. In the next chapter, Chapter 4, the most important reliability estimation procedures will be discussed more extensively.

The reliability of tests is, among others, influenced by test length, i.e. the number of parts or items in the test, and by the homogeneity of the group of subjects to whom the test is administered. This is the subject of Sections 3.4 and 3.5. Section 3.6 is concerned with the estimation of subject's true scores. Finally, we could ask ourselves what the correlation between two variables *X* and *Y* would be 'ideally', i.e. when errors of measurement affect neither variable. In Section 3.7 the correction for attenuation is presented.

## 3.2 The Definition of Reliability and the Standard Error of Measurement

An important development in the context of the classical true-score model is that of the concept of reliability. Starting from the variances and covariances of the components of the classical model, the concept of reliability can directly be defined. First, consider the covariance between observed scores and true scores. The covariance between observed and true scores, using the basic assumptions of the classical model discussed in Chapter 2, is as follows:

$$\sigma_{X\mathrm{T}} = \sigma(\mathrm{T}+E,\mathrm{T}) = \sigma^2_{\mathrm{T}} + \sigma(\mathrm{T},E) = \sigma^2_{\mathrm{T}}.$$

Now the formula for the correlation between true scores and observed scores can be derived as

$$\rho_{X\mathrm{T}} = \frac{\sigma_{X\mathrm{T}}}{\sigma_X\sigma_{\mathrm{T}}} = \frac{\sigma_{\mathrm{T}}}{\sigma_X},$$

the quantity also known as the reliability index. The reliability of a test is defined as the squared correlation between true scores and observed scores, which is equal to the ratio of true-score variance to observed-score variance:

$$\rho^2_{X\mathrm{T}} = \frac{\sigma^2_{\mathrm{T}}}{\sigma^2_X} = \frac{\sigma^2_{\mathrm{T}}}{\sigma^2_{\mathrm{T}}+\sigma^2_E}. \tag{3.1}$$

The reliability indicates to which extent observed-score differences reflect true-score differences. In many test applications it is important to be able to discriminate between persons, and a high test reliability is prerequisite. A measurement instrument that is reliable in a particular population of persons is not necessarily reliable in another population. From definition (3.1) it is clear that the size of the test reliability is population dependent. In a population with relatively small true-score differences reliability necessarily is relatively low.

Estimation of test reliability has always been one of the important issues in test theory. We will discuss reliability estimation extensively in the next chapter. For the moment we assume that reliability is known. Now we can define the concept of standard error of measurement. We derive from (3.1):

$$\sigma_{\mathrm{T}}^2 = \rho_{X\mathrm{T}}^2 \sigma_X^2 \tag{3.2}$$

and

$$\sigma_E^2 = \sigma_X^2 - \rho_{X\mathrm{T}}^2 \sigma_X^2 \, .$$

The standard error of measurement is defined as

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{X\mathrm{T}}^2} \, . \tag{3.3}$$

The reliability coefficient of a test and the standard error of measurement are essential characteristics (cf. *Standards*, APA, AERA, & NCME, 1999, Chapter 2). From the theoretical definition of reliability (3.1), and taking into account that variances cannot be negative, the upper and lower limit of the reliability coefficient can easily be derived as

$$0 \leq \rho_{X\mathrm{T}}^2 \leq 1.0$$

and $\rho_{X\mathrm{T}}^2 = 0$ if all observed-score variance equals error variance. If no errors of measurement occur, observed-score variance is equal to true score variance and the measurement instrument is perfectly reliable (assuming that there is true-score variation).

The observed-score variance is population or sample dependent, as is the reliability coefficient. So, only reporting the reliability coefficient of a test is insufficient, at least also the standard error of measurement must be reported.

## 3.3 The Definition of Parallel Tests

Loosely speaking, parallel tests are tests that are completely interchangeable. They are perfectly equivalent. But how to cast equivalence in statistical terms? Parallel tests are defined as tests that have identical true scores and identical person-specific error variances. Needless to say that parallel tests must measure the same construct or underlying trait.

For two parallel tests *X* and *X′* we have, as defined,

$$\tau_p = \tau'_p \text{ for all persons } p \text{ from the population} \tag{3.4a}$$

and

$$\sigma^2_{E_p} = \sigma^2_{E'_p} \text{ for all } p. \tag{3.4b}$$

Using the definition of parallel tests and the assumptions of the classical true-score model, we now can derive typical properties of two parallel tests $X$ and $X'$:

$$\mu_X = \mu_{X'}, \tag{3.5a}$$

$$\sigma^2_E = \sigma^2_{E'} \tag{3.5b}$$

$$\sigma^2_\mathrm{T} = \sigma^2_{\mathrm{T}'} \tag{3.5c}$$

$$\sigma^2_X = \sigma^2_{X'} \tag{3.5d}$$

and

$$\rho_{XY} = \rho_{X'Y} \text{ for all tests } Y \text{ different from tests } X \text{ and } X'. \tag{3.5e}$$

In other words, strictly parallel tests have equal means of observed scores, equal observed-score, true-score and error-score variances and equal correlations with any other test $Y$.

Now working out the correlation between two parallel tests $X$ and $X'$, it follows that

$$\rho_{XX'} = \frac{\sigma_{\mathrm{TT}'}}{\sigma_X \sigma_{X'}} = \frac{\sigma^2_\mathrm{T}}{\sigma^2_X} = \rho^2_{X\mathrm{T}}. \tag{3.6}$$

So, a second theoretical formulation of test reliability is that it is the correlation of a test with a parallel test. With this result we have obtained the first possibility to estimate test reliability: we can correlate the test with a parallel test. A critical note with this method, however, is how we should verify whether a second test is parallel. Also, parallelism is not a well-defined property: a test might have different sets of parallel tests (Guttman, 1953); see Exhibit 3.1. Further, if we do not have a parallel test, we must find another way to estimate reliability.

### Exhibit 3.1. On Parallelism and Other Types of Equivalence

To be sure, a certain test may have different sets of parallel tests (Guttman, 1953). Does it matter, for all practical purposes, if a test has different sets of parallel forms? An investigator will always look for meaningfulness and interpretability of the measurement results. If certain parallel forms do not suit the purpose of an investigator using a specific test, this investigator might well choose the most appropriate form of parallel test. Appropriateness may be checked against criteria relevant for the study at issue.

Parallel tests give rise to equal score means, equal observed-score and error means, and equal correlations with a third test. Gulliksen (1950) mentions the Votaw-Wiks' tests for this strict parallelism. These tests are also embedded in some computer programs for confirmatory factor analysis (a topic that is beyond the scope of our introductory text; see e.g. Byrne, 1994).

In Chapter 4 *congeneric* tests will be discussed, a form of equivalence of measures that moves away from the strict parallelism requirements. There are also other types of equivalence, less strict than parallelism, but stricter than equivalence according to the model of congeneric tests, which will be discussed in the next chapter.

## 3.4 Reliability and Test Length

In general, to obtain more precise measurements more observations of the same kind have to be collected. If we want a precise measure of body weight, we could increase the number of observations. Instead of one measurement, we could take 10 measurements, and take the mean of these observations. This mean is a more precise estimate of body weight than the result of a single measurement. This is what elementary statistics learns us. So, if we have a measurement instrument for which two or more parallel tests are available, we might consider the possibility to combine them into one longer, more reliable test. Assume that we have $k$ parallel tests. The variance of the true scores on the test lengthened by a factor $k$ is

$$\mathrm{var}(k\mathrm{T}) = k^2\sigma_{\mathrm{T}}^2 .$$

Due to the fact that the errors are uncorrelated the variance of the measurement errors of the lengthened test is

$$\mathrm{var}(E_1 + E_2 + \ldots . + E_k) = k\sigma_E^2 .$$

The variance of the measurement errors has a lower growth rate than the variance of true scores.

The reliability of the test lengthened by a factor $k$ is:

$$\rho_{X(k)X'(k)} = \frac{k^2\sigma_{\mathrm{T}}^2}{k^2\sigma_{\mathrm{T}}^2 + k\sigma_E^2} = \frac{k\sigma_{\mathrm{T}}^2}{k\sigma_{\mathrm{T}}^2 + \sigma_X^2 - \sigma_{\mathrm{T}}^2} .$$

After dividing numerator and denominator of the right hand side by $\sigma^2_X$ we obtain:

$$\rho_{X(k)X'(k)} = \frac{k\rho_{XX'}}{1 + (k-1)\rho_{XX'}} . \tag{3.7}$$

This formula is known as the general *Spearman-Brown formula for the reliability of a lengthened test*.

## 3.5 Reliability and Group Homogeneity

A reliability coefficient depends also on the variation of the true scores among subjects. So, the homogeneity of the group of subjects is an important characteristic to consider in the context of reliability. If a test has been developed to measure reading skill, then the true scores for a group of subjects consisting of children of a primary school will have a wider range, or a larger true-score variance, than the true scores of e.g. the 5-th grade children only. If we assume, as is frequently done, that the error score variance is equal for all relevant groups of subjects, we can compute the reliability coefficient for a target group from the reliability in the original group

$$\rho_{UU'} = 1 - \frac{\sigma_E^2}{\sigma_U^2} = 1 - \frac{\sigma_X^2(1-\rho_{XX'})}{\sigma_U^2} , \tag{3.8}$$

where $\sigma^2_U$ is the variance of the observed scores in the target group, $\sigma^2_X$ its counterpart in the original group and $\rho_{XX'}$ the reliability in the original group.

It is, however, advised to verify whether the size of the error variance varies systematically with the true score level. One method for the computation of the so-called conditional error variance, an important issue for reporting errors of measurement of test scores (see *Standards*, APA, AERA, & NCME, 1999, Chapter 2) has been suggested by Woodruff (1990). At several places in the book we will pay attention to the subject of conditional error variance.

## 3.6 Estimating the True Score

The true score can be estimated by the observed score and so it is done frequently. Assuming that the measurement errors are approximately normally distributed, we can construct a 95-percent confidence interval

$$x_p - 1.96\sigma_E \leq \tau_p \leq x_p + 1.96\sigma_E \,. \tag{3.9}$$

Unfortunately the point estimate and the confidence interval in (3.9) are misleading for two reasons. The first reason is that we can safely assume that the variance of measurement errors varies from person to person. Persons with a high or low true score have a relatively low error variance due to a ceiling and a floor effect, respectively. So, we should estimate error variance as a function of true score. We will return to this subject later on.

We will discuss the second reason in more detail. We start with a simple demonstration. Suppose all true scores are equal. Then the true score variance equals zero. So, the observed-score variance equals the variance of measurement errors. We know this because we have obtained a reliability equal to zero. Which estimate of a person's true score seems most adequate? In this case the best true-score estimate for all persons is the population mean $\mu_X$.

More generally, we might estimate $\tau$ using an equation of the form $ax_p + b$, where $a$ and $b$ are chosen in such a way that the sum of the squared differences between true scores $\tau$ and their estimates are minimal. The resulting formula is the formula for the regression of true score on observed score:

$$\tau = \frac{\sigma_{\mathrm{T}}\rho_{X\mathrm{T}}}{\sigma_X}(x - \mu_X) + \mu_{\mathrm{T}} \,.$$

This formula can be rewritten as:

$$\tau = \rho_{XX'}x + (1 - \rho_{XX'})\mu_X \,, \tag{3.10}$$

with a *standard error of estimation* (for estimating true score from observed score) equal to:

$$\sigma_\varepsilon = \sigma_{\mathrm{T}}\sqrt{1 - \rho^2_{X\mathrm{T}}} = \sigma_X\sqrt{\rho_{XX'}}\sqrt{1 - \rho_{XX'}} = \sqrt{\rho_{XX'}}\sigma_E \,. \tag{3.11}$$

Formula (3.10) is known as the Kelley regression formula (Kelley, 1947). From (3.11) it is clear that the Kelley estimate is better than the observed score as an estimate of true score.

The use of the Kelley formula can be criticized too:

1. The standard error of estimation (3.11) also supposes a constant error variance.
2. The true regression might be nonlinear.

3. The Kelley estimate of the true score depends on the population. Persons with the same observed score coming from different populations might have different true-score estimates and might consequently be treated differently.
4. The estimator is biased. The expected value of the Kelley formula equals $\tau_p$ only when the true score equals the population mean.
5. The regression formula is inaccurately estimated in small samples.

Under a few distributional assumptions the Kelley formula can be derived from a Bayesian point of view. Assume that we have a prior distribution of true score $\mathrm{N}(\mu_{\mathrm{T}}, \sigma_{\mathrm{T}}^2)$, i.e. the distribution is normal with mean $\mu_{\mathrm{T}}$ and variance $\sigma_{\mathrm{T}}^2$. Empirical-Bayesians take the estimated population distribution of T as the prior distribution of true scores. Also assume that the distribution of observed score given true score $\tau$ equals $\mathrm{N}(\tau, \sigma_E^2)$. Under these assumptions the mean of the posterior distribution of $\tau$ given observed score $x$ equals Kelley's estimate with $\mu_X$ replaced by $\mu_{\mathrm{T}}$. When a second measurement is taken, it is of course averaged with the first measurement in order to obtain a refined estimate of the true score. After a second measurement the variance of measurement errors is not equal to $\sigma_E^2$, but equal to $\sigma_E^2/2$. After $k$ measurements we have

$$\hat{\tau} = \frac{\sigma_{\mathrm{T}}^2}{\sigma_{\mathrm{T}}^2 + \sigma_E^2 / k} x(k) + \frac{\sigma_E^2 / k}{\sigma_{\mathrm{T}}^2 + \sigma_E^2 / k} \mu_{\mathrm{T}}, \tag{3.12}$$

where $x(k)$ is the average score after $k$ measurements, as the estimate of true score and as $k$ becomes larger, the expected value of (3.12) gets closer to the value $\tau$. So, the biasedness of the estimator does not seem to be a real issue.

## 3.7 Correction for Attenuation

The correlation between two variables $X$ and $Y$, $\rho_{XY}$, is small if the two true-score variables are weakly related. The correlation can also be small if one or both variables have a large measurement error. The correlation being weakened or attenuated due to measurement errors one might ask how large the correlation would be without errors, i.e. the correlation between the true-score variables. This is an old problem in test theory and the answer is simple. The correlation between the true-score variables is:

$$\rho_{\mathrm{T}_X \mathrm{T}_Y} = \frac{\sigma_{\mathrm{T}_X \mathrm{T}_Y}}{\sigma_{\mathrm{T}_X} \sigma_{\mathrm{T}_Y}} = \frac{\sigma_{XY}}{\sqrt{\rho_{XX'}}\sigma_X \sqrt{\rho_{YY'}}\sigma_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}}}. \tag{3.13}$$

The formula in (3.13) is the *correction for attenuation*. In practice the problem is to obtain a good estimate of reliability. Frequently only an underestimate of reliability is available. Then the corrected coefficient (3.13) can have a value larger than 1 in case the correlation between the true-score variables is high.

When data are available for several variables $X$, $Y$, $Z$ , etc. we can model the relationship between the latent variables underlying the observed variables. In Structural Equation Modeling the fit of the structure that has been proposed, can be investigated. So, structural equation modeling produces information on the true relationship between two variables.

## Exercises

3.1 The reliability of a test is 0.75. The standard deviation of observed scores is 10.0. Compute the standard error of measurement.

3.2 The reliability of a test is 0.5. Compute test reliability if the test is lengthened with a factor $k$ = 2, 3, 4, . . . , 16 ($k$ = 2(1)16 for short).

3.3 Compute the ratio of the standard error of estimation and the standard error of measurement for $\rho_{xx'} = 0.5$ and $\rho_{xx'} = 0.9$. Compute the Kelley estimate of true score for an observed score equal to 30, and $\mu_X = 40$, $\rho_{xx'} = 0.5$, respectively $\rho_{xx'} = 0.9$.

3.4 The reliability of test $X$ equals 0.49. What is the maximum correlation that can be obtained between test $X$ and a criterion? Explain your answer. Suggestion: use the formula for the correction for attenuation.

3.5 Let $\rho_{XY}$ be the validity of test $X$ with respect to test $Y$. Write the validity of test $X$ lengthened by a factor $k$, in terms of $\rho_{XY}$, $\sigma_X$, $\sigma_Y$ and $\rho_{XX'}$. What happens when $k$ becomes very large?

# 4 ESTIMATING RELIABILITY

## 4.1 Introduction

In the previous chapter reliability has been defined as the squared correlation between observed score and true score, which amounts to the ratio of true-score variance to observed-score variance. A nagging problem is how to tackle true-score variance, or, for that matter, error-score variance. Ultimately the latter question boils down to a succinct and unambiguous specification of the sources of measurement error at stake using the test. It must be noted that this is not entirely a question of statistical analysis, but also of logical and empirical analysis. In the present text the bulk of our attention will be paid to the statistical analysis of reliability.

In this chapter the major approaches to reliability estimation will be discussed. In the previous chapter we noticed that the test reliability is equal to the correlation between a test and a parallel test. The moment of administration of the second test, however, is of crucial importance as it may have an influence on error variance. If there is a long time interval between the administration of the first and the second test, the factor time may play an important role: persons may change in the time between testing. On the other hand, when tests are administered consecutively fatigue is to be expected to come into play. Therefore it is advisable with test administration in one session to split the persons into two groups, one of which is administered test $X$ first, followed by test $Y$, and the other is given the two test in reverse order.

As the *parallel-test method* is not without its problems (see again Guttman, 1953), an alternative method for reliability estimation would be to administer the test twice. This method is the *test-retest method.* With a small time interval between test sessions the risk is large that on the second test occasion persons remember their answers given on the first occasion. This would be a violation of the assumption of experimental independence. Clearly this violation would have a negative effect on the quality of the reliability estimate. With a larger time interval persons might be changed on the characteristic of interest. Therefore the test-retest method is only useful when a relatively stable characteristic is to be measured. The resulting reliability coefficient is called a stability coefficient for this reason.

There are also estimation methods based on data from a single administration of a test. These methods can be used when a test consists of several components, as most tests do. With these methods the momentary level of achievement of a respondent is taken as the true score of interest. Clearly a reliability coefficient obtained from the test at one occasion can differ from the stability coefficient. Table 4.1 gives an overview over the major approaches to reliability estimation.

In Section 4.2 estimation methods will be discussed based on a single administration of a test, in Section 4.3 methods with parallel tests and test-retest approaches, in Section 4.4 reliability and factor analysis, in Section 4.5 the estimation of true scores and score profiles, and in Section 4.6 the conditional standard error of measurement.

**TABLE 4.1.** Major approaches to reliability estimation

| | Reliability coefficient | Major error source | Data-gathering procedure | Statistical data-analysis |
|---|---|---|---|---|
| 1. | Stability coefficient (test-retest) | Changes over time | test-retest | product-moment correlation |
| 2. | Equivalence coefficient | item sampling from test form to test form | give form *j* and form *k* | product-moment correlation |
| 3. | Internal consistency coefficient | item sampling; test heterogeneity | a single administration | a) split-half correlation & Spearman-Brown correction<br>b) coefficient alpha<br>c) $\lambda_2$<br>d) other |

## 4.2 Reliability Estimation from a Single Administration of a Test

When a test is composed of several parts we might try to split the test into two parallel subtests. Then we might compute the correlation between the two halves. This correlation would give us an estimate of the reliability of a test with half the length of the original test. An estimate of the reliability of the original test can be obtained by applying the Spearman-Brown formula for a lengthened test. A weakness of the method is the arbitrary division of the test into two halves. Of course, this could easily be remedied by taking all possible splits into two halves. Should we confine ourselves to splits into two halves, however? The answer is no. Several coefficients have been proposed based on a split of a test into more than two parts (see Feldt & Brennan, 1989). We will discuss a method in which all parts or components play the same role.

Let test *X* be composed of *k* parts $X_i$. The observed score on the test can be written as

$$X = X_1 + X_2 + \ldots + X_k,$$

and the true score as

$$\mathrm{T} = \mathrm{T}_1 + \mathrm{T}_2 + \ldots + \mathrm{T}_k.$$

The reliability coefficient of the test is

$$\rho_{XX'} = \frac{\sigma^2_{\mathrm{T}}}{\sigma^2_X} = \frac{\sum_i^k \sigma^2_{\mathrm{T}_i} + \sum_i^k \sum_{j \neq i}^k \sigma_{\mathrm{T}_i \mathrm{T}_j}}{\sigma^2_X}.$$

The covariances between the true scores on the parts in the formula above equal the covariances between the observed scores on the parts. The true-score variances of the components are unknown. They can be approximated as follows.

While $(\sigma_{\mathrm{T}_i} - \sigma_{\mathrm{T}_j})^2 \geq 0$ we have

$$\sigma^2_{\mathrm{T}_i} + \sigma^2_{\mathrm{T}_j} \geq 2\sigma_{\mathrm{T}_i}\sigma_{\mathrm{T}_j}.$$

We also have

$\sigma_{\mathrm{T}_i}\sigma_{\mathrm{T}_j} \geq \sigma_{\mathrm{T}_i\mathrm{T}_j}$ , i.e. the correlation coefficient does not exceed 1,

so $(k-1)\sum_{i=1}^{k}\sigma^2_{\mathrm{T}_i} = \sum_{i<j}^{k-1}\sum_{j}^{k}(\sigma^2_{\mathrm{T}_i}+\sigma^2_{\mathrm{T}_j}) \geq \sum_{i=1}^{k}\sum_{j\neq i}^{k}\sigma_{\mathrm{T}_i\mathrm{T}_j}$ .

From this we obtain:

$$\rho_{XX'} = \frac{\sum_{i=1}^{k}\sigma^2_{\mathrm{T}_i} + \sum_{i=1}^{k}\sum_{j\neq i}^{k}\sigma_{\mathrm{T}_i\mathrm{T}_j}}{\sigma^2_X} \geq \frac{\frac{k}{k-1}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\sigma_{\mathrm{T}_i\mathrm{T}_j}}{\sigma^2_X} = \frac{\frac{k}{k-1}\sum_{i=1}^{k}\sum_{j\neq i}^{k}\sigma_{X_iX_j}}{\sigma^2_X} = \left(\frac{k}{k-1}\right)\frac{\sigma^2_X - \sum_{i=1}^{k}\sigma^2_{X_i}}{\sigma^2_X} .$$

Now we have obtained a lower bound to the reliability. The coefficient is referred to as coefficient α:

$$\alpha \equiv \left(\frac{k}{k-1}\right)\left(1-\frac{\sum_{i=1}^{k}\sigma^2_{X_i}}{\sigma^2_X}\right). \tag{4.1}$$

Coefficient α is also called a measure for internal consistency. We can elucidate the reason for this designation with an example. Let us take an anxiety questionnaire. Let us assume that different persons experience anxiety in different situations. Test reliability as estimated by coefficient α might be low although anxiety might be a stable characteristic. The test-retest method might have given a much higher reliability estimate.

The popularity of coefficient α is due to Cronbach (1951). The coefficient was proposed earlier by Hoyt (1941) on basis of an analysis of variance (see Chapter 5), and by Guttman (1945) as one of a series of lower bounds to reliability. Therefore McDonald (1999, p. 95) refers to this coefficient as Guttman-Cronbach alpha. Following the *Standards* (APA, AERA, & NCME, 1999), however, we shall stick to the use of calling it Cronbach's alpha. For dichotomous items the item variance of item $i$ can be simplified to $p_i(1 - p_i)$ if we divide by the number of persons $N$ in the computation of the variances instead of $N$ - 1. Here $p_i$ is the proportion of correct responses to the item. The resulting coefficient is called Kuder-Richardson formula 20, KR20 for short (Kuder & Richardson, 1937). Kuder and Richardson also proposed a further simplification, KR21. In KR21 all $p_i$ are replaced by the average proportion correct. When the item difficulties are unequal, KR21 is lower than KR20. KR21 is discussed later on in the chapter on dichotomous items.

Under certain conditions coefficient α is not a lower bound to reliability but an estimate of reliability itself. This is the case if all items (components) have the same true-score variance and if the true scores of the items correlate perfectly. In this case the two inequalities in the derivation of the coefficient become equalities. Items or tests that satisfy this property are called (essentially) tau-equivalent. The definition of *essentially tau-equivalent tests i* and *j* is:

$$\tau_{ip} = \tau_{jp} + b_{ij}. \tag{4.2}$$

If true scores are equal, i.e. if the additive constant $b_{ij}$ equals 0; we have *tau-equivalent measurements*. Tau-equivalent tests with unequal error variances have unequal reliabilities. If true scores and error variances are equal, we have parallel tests. In the case of parallel test items coefficient $\alpha$ can be rewritten in the form of the Spearman-Brown formula for the reliability of a lengthened test (3.7) where the reliability at the right hand side of the '='-sign in the formula is replaced by the common intercorrelation between items.

A further relaxation of (4.2) would be if the true score of tests *i* and *j* are linearly related, i.e. if

$$\tau_{ip} = a_{ij}\tau_{jp} + b_{ij}. \tag{4.3}$$

In this case we have the model of *congeneric tests*: true-score variances, error variances as well as population means can be different. The congeneric test model is the furthest relaxation of the classical test model.

Let us have a further look at (4.3). In (4.3) the true score on test *i* is defined in terms of the true score on test *j*. An alternative and preferable formulation would be to write true scores on test *i* as well as test *j* in terms of a latent variable T. So,

$$\tau_{ip} = a_i\tau_p + b_{i,} \tag{4.4a}$$

and

$$\tau_{jp} = a_j\tau_p + b_j. \tag{4.4b}$$

The true-score variances are $a_i^2\sigma^2_T$ and $a_j^2\sigma^2_T$. Without loss of generality we can set $\sigma^2_T$ equal to 1. For, if $\sigma_T$ has a value $u$ unequal to 1, we can define a new latent score $\tau^*$ and new coefficients $a^*$ with

$$\tau^* = \tau/u \text{ and } a^* = a \times u,$$

and the new latent score has a variance equal to 1. The variances of the congeneric tests can be written as

$$\sigma_i^2 = a_i^2 + \sigma_{E_i}^2 \tag{4.5}$$

and the covariances as

$$\sigma_{ij} = a_i a_j\,. \tag{4.6}$$

With three congeneric tests there are three observed-score variances and three different covariances. There are six unknown parameters: three coefficients *a* and three error variances. The unknown parameters can be computed from the observed-score variances and covariances. With two congeneric tests we have more unknowns than observed variances and covariances. In this case we cannot estimate the coefficients *a* and the error variances. With more than three tests more variances and covariances are available than unknown parameters. Then a statistical estimation procedure is needed in order to estimate the parameters from the data according to a specified criterion. Such a procedure is computer implemented in software for structural equation modeling (LISREL, EQS, AMOS).

It is important to have more than three tests when the congeneric test assumption is to be verified (three tests are enough to verify whether the stronger assumption of parallelism is satisfied). The advantage of the exact computation of the coefficients *a* and the error variances in the case of three tests is only apparent. For, even when tests are not congeneric it is possible to compute three values *a* for three tests and in most cases also realistic error variances (with nonnegative values) are obtained. With more than three tests the assumption that tests are congeneric can be tested (Jöreskog, 1971). If the congeneric test model fits, we can also verify whether a more restrictive model - the (essentially) tau-equivalent test model or the model with parallel tests – fits the data. If a simpler, more restrictive model adequately fits the data, this model is to be preferred. It is also possible that the congeneric model does not fit. Then we can try to fit a structural model with more than one dimension (Byrne, 1994; Jöreskog & Sörbom, 1993).

The administration of a number of congeneric tests is practically unfeasible. However, an existing test might be composed of subtests that are congeneric, tau-equivalent or even parallel. In such a situation the method for estimating coefficients *a* for congeneric measurements can be used for the estimation of test reliability. If we have congeneric subtests, the estimate of reliability is

$$\rho_{XX'} = \frac{\left(\sum_{i=1}^{k} a_i\right)^2}{\sigma_X^2} . \tag{4.7}$$

If coefficients *a* and error variances of the subtests are available, it is possible to use them for computing weights that maximize reliability. Jöreskog (see also Overall, 1965) demonstrated that with congeneric measurements optimal weights are proportional to

$$w_i = \frac{a_i}{\sigma_{E\,i}^2} . \tag{4.8}$$

In other words, the optimal weight is smaller for a large error variance and higher in case the subtest contributes more to the true score of the total test. More information on weighting is given in Exhibit 4.1.

Let us now return to coefficient $\alpha$ and the question of alpha as a lower bound to the reliability of a test. For a test composed of a reasonably large number of items that are not too heterogeneous, coefficient $\alpha$ underestimates reliability slightly. On the other hand, it is possible for coefficient $\alpha$ to have a negative value, although reliability itself – being defined as the ratio of two variances - cannot be negative. Better lower bounds than coefficient $\alpha$ are available. Guttman (1945) derived several lower bounds. One of these, called $\lambda_2$, is always equal to or larger than coefficient $\alpha$. The formula for this coefficient is

**Exhibit 4.1. Weighting responses and variables.**

A total score is obtained by adding item scores. The total score can be an unweighted sum of the item scores or a weighted sum score. Two kinds of weights are in use: a priori weights and empirical weights. Empirical weights are somehow based on data. Many proposals for weighting have been done. Among these proposals are the optimal weights for congeneric measurements and weights that are defined within the context of Item Response Theory.

We mention one other proposal for weights here: the weighting of item categories and items on basis of a homogeneity analysis. Homogeneity analysis, or dual scaling , is used for scaling variables that are defined on an ordinal scale. Weights are assigned to the categories of these variables. The weights and scores have symmetrical roles. A person’s score is defined as the average of the category weights of the categories that have been endorsed. The category weight of a variable is proportional to the average score of the persons who have chosen the category. Actually, one of the algorithms to obtain weights and scores is to iterate between computing scores on basis of the weights and weights on the basis of the scores until convergence has been reached.

Lord (1958) has demonstrated that the category weights of homogeneity analysis maximize coefficient alpha. With dichotomous items item weights can be defined instead of separate weights for the two categories. Further information on the subject can be found in Nishisato (1980). A more readable introduction to dual scaling (or optimal scaling or correspondence analysis) is Nishisato (1984).

The so-called maxalpha weights are optimal weights within the context of homogeneity analysis. In other approaches other weights are found to be optimal. A general treatment of weighting is given by McDonald (1968). When items are congeneric the weights that maximize reliability are obviously optimal, and these weights are not identical to the maxalpha weights. The ultimate practical question is: is differential weighting of responses and variables worth the trouble. In the context of classical test theory the answer is ‘seldom’. Usually, items are selected that are highly correlated. Then the practical significance is limited (cf. Gifi, 1990, p. 84).

$$\lambda_2 \equiv \frac{\sigma_X^2 - \sum_{i=1}^{k} \sigma_{X_i}^2 + k\sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i}^{k} \sigma_{X_i X_j}^2}}{\sigma_X^2} . \tag{4.9}$$

In Exhibit 4.2 an example of reliability estimation with a number of coefficients is presented.

Still two other lower bounds are worthwhile to discuss. The first one is the g.l.b., the ‘greatest lower bound’; its definition will be discussed in Section 4.4. The second one is coefficient $\alpha_s$, the stratified coefficient $\alpha$ (Rajaratnam, Cronbach & Gleser, 1965).

First, let us rewrite coefficient $\alpha$ as

$$\alpha = \frac{k^2 \text{ave}(\sigma_{ij})}{\sigma_X^2} , \tag{4.10}$$

where *ave* denotes average and $\sigma_{ij}$ is shorthand for the covariance between item *i* and item *j*. Figure 4.1 illustrates the situation for a 4-item test. The diagonal entries in the figure represent the item variances. The off-diagonal entries represent the covariances between items. The sum of the entries equals the variance of the total test, the denominator in (4.10). The numerator of coefficient $\alpha$ according to (4.10) is obtained by replacing all diagonal values in the figure by the average covariance and, next, summing all entries.

### Exhibit 4.2. An Example with Several Reliability Estimates

Lord and Novick (1968, p. 91) present the variance-covariance matrix for four components, based on data for the Test of English as a Foreign Language. Their data are replicated in the table below. From the table we can read that the variance of the first component equals 94.7; the covariance between components 1 and 2 equals 87.3.

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| C1 | 94.7 | 87.3 | 63.9 | 58.4 |
| C2 | 87.3 | 212.0 | 138.7 | 128.2 |
| C3 | 63.9 | 138.7 | 160.5 | 109.8 |
| C4 | 58.4 | 128.2 | 109.8 | 115.8 |

We use the data in the table for the computation of several reliability coefficients. First, let us compute split-half coefficients with a Spearman-Brown correction for test length. The total test can be split into two half test in three different ways. We compute all three possible reliability estimates.

| split(a,b) | var(a) | var(b) | cov(a,b) | r | r(2) |
|---|---|---|---|---|---|
| 12-34 | 481.30 | 495.90 | 389.20 | 0.797 | 0.887 |
| 13-24 | 383.00 | 584.20 | 394.20 | 0.833 | 0.909 |
| 14-23 | 327.30 | 649.90 | 389.20 | 0.844 | 0.915 |

The estimates vary from 0.887 to 0.915. An alternative approach on basis of the split of the test into two halves, would have been to use coefficient alpha with two components.

Next we compute coefficient alpha. The total score variance is equal to the sum of all cell values in the table: 1755.6. The sum of the component variances equals: 583.0. Coefficient alpha equals

$$\alpha = (4/3)\ (1 - 583.0/1755.6) = 0.891.$$

The value of $\alpha$ is lower than the highest estimate based on a split into two parts. Coefficient alpha is guaranteed a lower bound to reliability, the split-half coefficient not. The most adequate estimate based on splitting the test in halves seems to be the first, because the split 12-34 seems to produce more or less comparable halves.

Finally, we compute $\lambda_2$. Therefore we need the square root of the average value of the squared covariances: 102.3426. We obtain:

$$\lambda_2 = (1755.6 - 583.0 + 4 \times 102.3426)/1755.6 = 0.901.$$

The value of $\lambda_2$ is higher than the value of $\alpha$.

| | | | |
|---|---|---|---|
| $\sigma^2_1$ | $\sigma_{12}$ | $\sigma_{13}$ | $\sigma_{14}$ |
| $\sigma_{21}$ | $\sigma^2_2$ | $\sigma_{23}$ | $\sigma_{24}$ |
| $\sigma_{31}$ | $\sigma_{32}$ | $\sigma^2_3$ | $\sigma_{34}$ |
| $\sigma_{41}$ | $\sigma_{42}$ | $\sigma_{43}$ | $\sigma^2_4$ |

**FIGURE 4.1.** The variance-covariance matrix for a 4-item test

Now suppose that we can classify the items into two relatively homogeneous clusters or strata. We can use this stratification in the computation of the estimated total true-score variance. We can replace the item variances within a stratum by the average covariance between items belonging to this stratum instead of by the average covariance computed over all item pairs. So, in the example in Figure 4.2 the variances of item 1 and 2 are replaced by $\sigma_{12}$ (= $\sigma_{21}$).

$$\begin{array}{c|cc|cc|}
 & \multicolumn{2}{c}{\text{stratum 1}} & \multicolumn{2}{c|}{\text{stratum 2}} \\
\hline
\text{stratum 1} & \sigma^2_1 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\
 & \sigma_{21} & \sigma^2_2 & \sigma_{23} & \sigma_{24} \\
\text{stratum 2} & \sigma_{31} & \sigma_{32} & \sigma^2_3 & \sigma_{34} \\
 & \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma^2_4 \\
\hline
\end{array}$$

**FIGURE 4.2.** The variance-covariance matrix of a 4-item test with two strata

The stratified coefficient can be written as

$$\alpha_s = \frac{\sum_{i=1}^{q} \alpha(i)\sigma^2_{Y_i} + \sum_{i=1}^{q}\sum_{j \neq i}^{q} \sigma_{Y_i Y_j}}{\sigma^2_X}, \tag{4.11}$$

where $q$ is the number of strata, $Y_i$ the observed score in stratum $i$ and $\alpha(i)$ coefficient $\alpha$ computed over the items in stratum $i$. A more general reliability formula from test theory is obtained if we replace $\alpha(i)$ in (4.11) by a possibly different reliability estimate for subtest $i$.

Reliability estimation based on a measure of internal consistency is problematic in case the item responses cannot be considered experimentally independent. This might happen, for example, if the test is answered under a time limit and some persons do not reach the items at the end of the test.

We always estimate reliability in a sample from the population of interest. With a small sample we must be alert to the risk that the reliability estimate in the sample deviates notable from the value in the population. An impression of the extent to which the sample estimates might vary can be obtained by splitting the sample into two halves and computing the reliability coefficient in both (a procedure that gives an impression of the variability in samples half the size of the sample in the investigation). We also can obtain an estimated sampling distribution on basis of some distributional assumptions. Distributional results for coefficient $\alpha$ can be found in Pandey & Hubert (1975) among others. One might also obtain sampling results with the bootstrap (Efron & Tibshirani, 1993). Raykov (1998) reports a study using the bootstrap for obtaining the standard error for a reliability coefficient.

## 4.3 Reliability Estimation with Parallel Tests and Test-Retest Approaches

Interchangeable test forms in terms of the definition of parallelism are used for the parallel-forms or alternate-forms of reliability. It has been noticed that parallel tests are not uniquely defined. The same test could belong to more than one set of parallel tests, leading, in general, to more than one reliability coefficient. Often, however, e.g. when dealing with the reliability of pure speed tests and partly speeded tests, the parallel-forms approach is the only feasible one. The best what one could do if the parallel-forms approach is followed, is to specify the conditions of the specified parallel measurement context.

A practical problem of the parallel-forms approach is that the instruments might not satisfy the requirements of parallel tests. Then alternative equivalence models as given in Section 4.2 could be considered. The reliability could be estimated from the results of an analysis with software for structural equation modeling.

The test-retest approach is appropriate if one may reasonably expect the construct or trait to be measured, to stay relatively stable within the period of measurement, and if memory effects do not influence the responses on the retest. It is obvious that calculating the correlation coefficient gives the reliability estimate.

## 4.4 Reliability and Factor Analysis

The analysis with congeneric measurements in Section 4.2 is an example of a linear factor analysis, to be precise a factor analysis with one common factor and as many specific and error factors as there are congeneric measurements. One hopes for one dominant common factor, indicating that one is measuring a single construct. Frequently, a factor analysis results in more than one common factor, however. Assume that we have factor analyzed a number of subtests and have found $m$ common factors. Then the score of person $p$ on subtest $i$ can be written as

$$x_{ip} = \mu_i + a_{i1}\theta_{1p} + a_{i2}\theta_{2p} + \ldots + a_{im}\theta_{mp} + e_{ip}, \tag{4.12}$$

where the $\theta'$s are common factors with mean 0 and variance 1 and $e_i$ is the measurement error. In (4.12) no subtest-specific factor has been defined. Without replicate measurements for a subtest the specific factor for this subtest can be regarded as part of the measurement error. So the true score from test theory is written as the weighted sum of common factors 1 through $m$ plus an overall mean. In Exhibit 4.3 a graphical example of the true score and factor representation of (sub)tests is given.

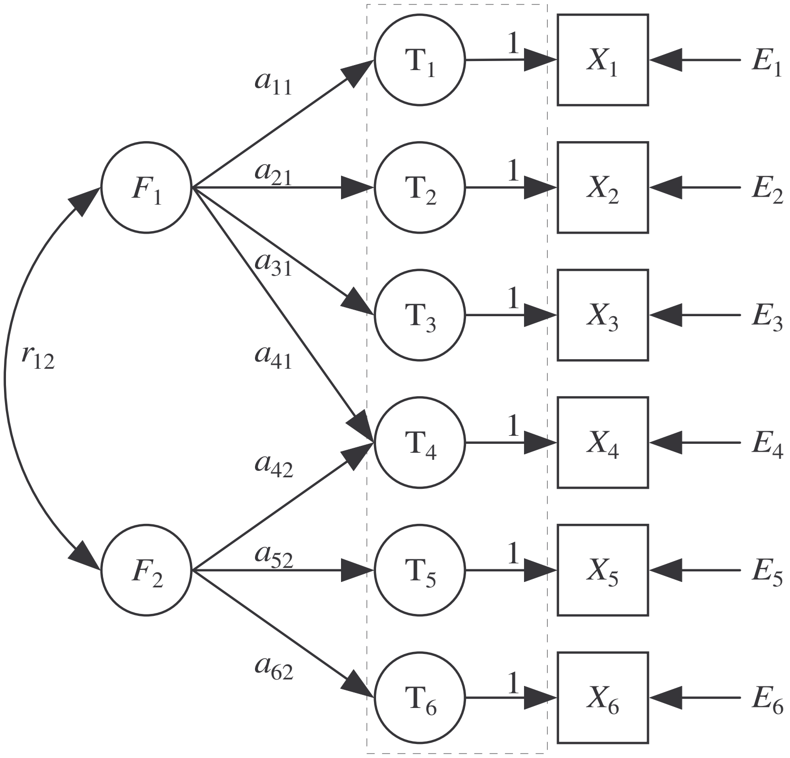
In a factor analysis several choices have to be made that influence the sizes of the estimated error variances and, consequently, the size of the reliability coefficient that can be computed.

The greatest lower bound to reliability (g.l.b.) can be obtained from a special kind of factor analysis. The determination of the g.l.b. is a complicated problem (Ten Berge, Snijders & Zeegers, 1981).

**Exhibit 4.3. True Scores and the Dimensionality of Tests**

In the structural equation modeling approach we can define a specific factor model and estimate the parameters of this model. In the figure below a two-factor model with correlated factors is presented in the graphical notation of the structural modeling approach. The first three (sub)tests load only on the first factor, (sub)tests 5 and 6 load on factor 2, and (sub)test 4 loads on both factors.

For illustrative purposes the representation of the true scores is added to the figure (in the dashed box).



From the figure we obtain

$$\sigma_4^2 = a_{41}^2 + a_{42}^2 + 2a_{41}a_{42}r_{12} + \sigma_{E_4}^2 ,$$

$$\sigma_{14} = a_{11}a_{41} + a_{11}a_{42}r_{12} ,$$

etc.

## 4.5 Score Profiles and Estimation of True Scores

Factor analysis frequently is used in test construction. We might construct a large number of items thought to be relevant to the characteristic to be measured. In an exploratory factor analysis we can verify whether the items are factorially homogeneous. If this is the case, we have a one-dimensional construct. It is, however, also possible to obtain more dimensions from an analysis. In the latter case the factor analysis also gives information on the possibility to cluster items into groups measuring various dimensions of interest. In a confirmatory analysis we hypothesize a priori a factor structure and verify whether the data support the hypothesis.

If several subtests are defined we can compute a total score as well as a separate score for each of the subtests. The higher the correlation between subtests and the less reliable subtests are, the less useful it is to compute subtest scores besides a total score. With reliable subtests which do not correlate too strongly, it makes sense to compute subtest scores besides or instead of total test score.

With subtest scores we can compute a score profile for each person tested. We can verify whether a person has relatively strong and relatively weak points. We can determine to which extent the score profile of a person is deviant. For a solid interpretation of the profile of scores it is important to standardize the subtests so that they have the same score distribution in the relevant population of persons. At least, the subtests should have identical means and standard deviations (for norms with respect to the estimation of means see Angoff, 1971; for sampling techniques, see Kish, 1987). Only then it is relatively simple to notice whether a person scores relatively high on one subtest and relatively low on another, for example, relatively high on a verbal subtest and relatively low on a mathematical subtest. When subtest reliabilities vary notably the advantage of this way of scaling the subtests is limited, however, for then there are large differences between the true-score scales (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

The observed score on a subtest is an obvious estimate of the true score on the particular subtest. In a previous section it was demonstrated that the observed score was not optimal: the Kelley estimate performs better than the observed score. For profile scores one can think of generalizing Kelley's formula.

Let us take a profile with two subtests *X* and *Y* as an example. We are interested in the true score of a person *p* on subtest *X*, $\tau_p(X)$. If we knew the true scores on test *X*, we would certainly consider the possibility to 'predict' these scores from observed scores *x* and *y* using multiple regression. There is no reason not to use the multiple regression formula in case the criterion $\tau$ is unknown. The formula with which we 'predict' the true score on subtest *X* is

$$\hat{\tau}_p(X) = \mu_{\mathrm{T}(X)} + \frac{\sigma_{\mathrm{T}(X)}}{\sigma_X}\frac{\rho_{\mathrm{T}(X)X} - \rho_{\mathrm{T}(X)Y}\rho_{XY}}{1-\rho_{XY}^2}(x_p - \mu_X) + \frac{\sigma_{\mathrm{T}(X)}}{\sigma_Y}\frac{\rho_{\mathrm{T}(X)Y} - \rho_{\mathrm{T}(X)X}\rho_{XY}}{1-\rho_{XY}^2}(y_p - \mu_Y)\,. \quad (4.13)$$

In this formula several correlations with true scores on *X* are involved and these correlations are unknown. Also unknown is the standard deviation of true scores on subtest X. However all the unknowns can be estimated:

$$\sigma_{\mathrm{T}(X)} = \sqrt{\rho_{XX'}}\sigma_X\,,$$

$$\rho_{\mathrm{T}(X)X} = \sqrt{\rho_{XX'}}$$

and, analogously to the correction for attenuation for two variables *X* and *Y*,

$$\rho_{\mathrm{T}(X)Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}}}\,.$$

We can conclude that the best estimate (in the least squares sense) of the true score on subtest *X* makes use of the score on subtest *Y* as well. With reliable *X* the score on subtest *X* gets a high weight. The weight of test *X* is also relatively high in case the scores on subtest *Y* are nearly uncorrelated with those on subtest *X*. In case true scores (and observed scores) on test *Y* are uncorrelated with those on *X*, the formula can be simplified to the Kelley formula. With congeneric subtests *X* and *Y* the obtained weights equal the optimal weights for congeneric measurements (4.8). In case true scores on subtests *X* and *Y* are strongly correlated and subtest *X* is relatively unreliable, it is possible to have a smaller weight for *X* than for *Y* in the formula for the prediction of the true score on *X*.

It is instructive to write Equation (4.13) in terms of variances and covariances:

$$\hat{\tau}_p(X) = \mu_X + \frac{\sigma_{\mathrm{T}(X)X}\sigma_Y^2 - \sigma_{\mathrm{T}(X)Y}\sigma_{XY}}{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}(x_p - \mu_X) + \frac{\sigma_{\mathrm{T}(X)Y}\sigma_X^2 - \sigma_{\mathrm{T}(X)X}\sigma_{XY}}{\sigma_X^2\sigma_Y^2 - \sigma_{XY}^2}(y_p - \mu_Y)\,. \qquad (4.14)$$

We can rewrite this equation as follows:

$$\hat{\tau}_p(X) = \hat{\tau}_p(X) \mid y_p + \frac{\sigma_{\mathrm{T}(X)}^2 - \sigma_{\mathrm{T}(X)Y}^2/\sigma_Y^2}{\sigma_X^2 - \sigma_{\mathrm{T}(X)Y}^2/\sigma_Y^2}(x_p - \hat{\tau}_p(X) \mid y_p)\ , \qquad (4.15)$$

where

$$\hat{\tau}_p(X) \mid y_p = \mu_X + \frac{\sigma_{\mathrm{T}(X)Y}}{\sigma_Y^2}(y_p - \mu_Y)\,. \qquad (4.16)$$

In other words, the optimal prediction formula for predicting the true score on *X* given observed scores on *X* and *Y* can be viewed as a two-step process. First, we estimate the true score on *X* given the observed score on *Y*. Next, we improve this estimate using the information given by the observed score on *X*. This way of viewing the estimation procedure would be quite logical if we take different measurements at different occasions. For example, *Y* might be the first measurement with a measurement instrument and *X* the second. In the Kalman filter the estimate of true score on time *t* is based on test data obtained at time *t* and the true score estimate at time *t* – 1. (Oud, Van den Bercken & Essers, 1990).

The estimation of profile scores with Equation (4.13) can evoke similar objections as application of Kelley’s formula in connection with a single test. The estimate of a person’s true score depends on the population that serves as a reference. Certainly, when we use profile scores, it is obvious that we compare the outcomes for a person with results for a reference population whether we use Kelley’s formula or not. The subtests are scaled in such a way that they have the same mean in some population. And when more than one relevant population exists, there is nothing against making separate norms for these different populations.

Another disadvantage of the use of a formula like (4.13) seems to be the detection of persons with deviant score patterns. Let two subtests correlate strongly. The estimation formula then gives similar estimates of the two true scores. The relevant information that the pattern of scores is deviant is likely to be missed.

We can find out whether the score pattern is aberrant. We will demonstrate this with observed scores on two tests *X* and *Y*. The prediction of the score on test Y, given the score on test *X*, is given by the regression equation

$$\hat{Y} = \sigma_Y \rho_{XY}(X - \mu_X)/\sigma_X + \mu_Y\,, \qquad (4.17)$$

with a standard error of prediction equal to

$$\sigma_\varepsilon = \sigma_Y\sqrt{1 - \rho_{XY}^2}\ . \qquad (4.18)$$

We can compute the predicted value on test $Y$ and construct a 95 percent confidence interval using the assumption of normally distributed prediction errors. If the observed score on test $Y$ lies outside this interval, we have an argument to consider the score pattern as aberrant. Of course we might also evaluate the raw score difference $x - y$. Then we evaluate the difference irrespective of the correlation between $X$ and $Y$. The relevant standard deviation for the raw-score difference is

$$\sigma_{E(X\text{-}Y)} = \sqrt{\sigma^2_{E(X)} + \sigma^2_{E(Y)}} = \sqrt{\sigma^2_X(1-\rho_{XX'}) + \sigma^2_Y(1-\rho_{YY'})}\,. \tag{4.19}$$

With (4.19) we can construct a 95 percent confidence interval for the difference between true scores on tests $X$ and $Y$. When the true scores on tests $X$ and $Y$ have a correlation smaller than 1, then more, perhaps much more than 5 percent of the observed differences falls outside of this interval.

A special application of profiles is that in which scores $X$ and $Y$ are two measurements on the same measurement instrument, taken at two different occasions. Now we might be interested in the possibility of a true-score change or a true-score gain. The simplest way to estimate the true difference is to use the difference score. However, difference scores have a bad reputation. They can be quite unreliable even in case the separate measurements are highly reliable. Difference scores are used when the two measurements are related. So, we may assume that the true scores on both measurements are strongly correlated. Let us suppose that we have a situation in which the true scores on both measurements are equal. Then the true change is zero and the reliability of difference scores is zero too.

On the other hand, a low reliability does not imply that there are no changes. It is possible that all persons have changed the same amount between testing occasions. On the group level the measurement of change is useful even with a low reliability for difference scores.

The presence of measurement error affects change in a special way. Let us analyze this in a simple situation in which the mean and variance of scores are equal in a pretest and a posttest. We will notice that there are changes although there is no overall change. The persons with better than average scores on the pretest will on the average have lower scores on the posttest. Persons with lower than average scores on the pretest will on the average show some score gain. The scores regress to the mean. This effect appears even if there is no true change at all. The effect is due to measurement error. Among the high scores on the pretest there always are some scores which are high due to measurement error. Among the low scores on the pretest there always are scores low due to measurement error. The difference score itself (posttest – pretest) is negatively correlated with the measurement error on the pretest. The true difference better is estimated by equations like (4.13) (Lord & Novick, 1968, pp. 74-6).

Due to the regression to the mean the use of difference scores in research is problematic. For research, alternatives are available (see Burr & Nesselroade, 1990; Cronbach & Furby, 1970). Rogosa, Brandt & Zimowski (1982) discuss the possibility of modeling growth in case of more than two occasions.

## 4.6 Reliability and Conditional Errors of Measurement

The 1985 as well as the 1999 *Standards* (APA, AERA, & NCME) emphasized to report reliability as well as the standard errors of measurement. And, in addition, Standard 2.2 of the 1999-edition states:

> The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw scores or original scale units and in units of each derived score recommended for use in test interpretations. (*Standards*, APA, AERA, & NCME, 1999, p.31).

The standard error of measurement can vary with true score level. Conditional standard errors of measurement are standard errors of measurement conditional on true score level. Such standard errors of measurement can be used as an alternative approach to convey reliability information, by constructing a confidence interval for an examinee's true score, universe score (to be discussed in Chapter 5) or percentile rank. Earlier three types of standard errors have been discussed: the standard error of measurement (Equation 3.3), the standard error of estimate of true score (Equation 3.11) and the standard error of prediction (equation 4.18).

Woodruff (1990) studied the conditional standard error of measurement for assessing the precision of a test on its score scale. He proposed to split a test into two parallel halves $X$ and $X'$. ANOVA is used to estimate values $\sigma^2(E'|X)$ as substitutes of $\sigma^2(E'|\mathrm{T})$. Then the outcomes are corrected for the fact that the test was split into two halves (using the customary assumption that the error variance doubles for a test lengthened by a factor 2).

Feldt & Qualls (1996) proposed a method for the estimation of the conditional error variance based on a split of the test into a number of essentially tau-equivalent subtests. It is possible to use a split of the test into two halves , but it proves to be better to split the test in many subtests as long as all subtests can be considered as essentially tau-equivalent measurement instruments. Let there be $n$ subtests. For person $p$ the estimated error variance of the subtests is

$$s^2_{E(p)} = \frac{\sum_{i=1}^{n}[(x_{pi} - x_{p.}) - (x_{.i} - x_{..})]^2}{n-1}, \tag{4.20}$$

where the scores are corrected for the test effects $x_{.i}$ - $x_{..}$. In the terminology of ANOVA two-way interactions are used in (4.20). Suppose that the subtests have equal score ranges. Then the consequence of the assumption of essentially tau-equivalent subtests on which Equation (4.20) is based, is that the error variance associated with a perfect score is nonzero when the subtests differ in difficulty level. In a nonlinear true-score model, a model based on IRT, such a strange effect does not occur.

Again, we must multiply the estimate with a constant in order to obtain the error variance on the total test. When the $n$ subtests add up to the total test, total test length is $n$ times the length of the subtests and the result in (4.20) must be multiplied by $n$.

Next, the error variances for all persons with the same total score can be averaged. This produces the estimated relationship between the size of the conditional error variance and total score. Feldt and Qualls suggest to reduce sampling variation further by smoothing the empirical relationship between error variance and total score. This can be achieved by a polynomial regression, where the error variance is regressed on powers of $X$ ($X$, $X^2$, etc.).

It might be interesting to compare (4.20) with a formula for the conditional error variance developed within the context of generalizability theory. For this purpose (4.20) is rewritten as

$$s^2_{E(p)} = s^2(x_{pi} \mid p) + s^2(x_{.i}) - 2\,\mathrm{cov}(x_{pi}, x_i \mid p)\,, \tag{4.21}$$

which is comparable to (5.40).

More methods for estimating conditional standard errors of measurement are described by Lee, Brennan & Kolen (2000). Methods for obtaining conditional error variances have been proposed specifically within the context of generalizability theory (Chapter 5) and for dichotomous items (Chapter 6). In item response theory the problem of the conditional standard error of measurement is approached in another way (see Section 6.4).

## Exercises

4.1 A test $X$ is given with three subtests, $X_1$, $X_2$ and $X_3$. The variance-covariance matrix for the subtests is given in the table below. Estimate reliability with coefficient $\alpha$.

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $X_1$ | 8.0 | 6.0 | 8.0 |
| $X_2$ | 6.0 | 12.0 | 12.0 |
| $X_3$ | 8.0 | 12.0 | 17.0 |

4.2 Use the variance-covariance matrix from exercise 4.1 for estimating test reliability according to the model of congeneric tests. Use (4.6) for the estimation of the $a_i$.

4.3 Prove that for parallel test items coefficient alpha equals the Spearman-Brown formula for the reliability of a lengthened test.

4.4 Two tests $X_1$ and $X_2$ are congeneric measurement instruments. Their correlations with other variables $Y_1$, $Y_2$, . . differ. Is there a pattern to be found in the correlations?

4.5 Given are two tests $X$ and $Y$ with $\sigma^2_X = 16.0$, $\sigma^2_Y = 16.0$, $\rho_{XX'} = \rho_{YY'} = 0.8$ and $\rho_{XY} = 0.7$.

a. Compute the observed-score variance, the true-score variance and the reliability of the difference scores $X - Y$.

b. Compare the variance of the raw score differences with $\sigma^2_{E(X-Y)}$ of (4.19).

4.6 In a test several items cover the same subject. Which assumption of classical test theory might be violated? What should we do when we want to estimate reliability with coefficient $\alpha$?

4.7 We have three tests $X_1$, $X_2$ and $X_3$ measuring the same construct. Their correlations with test $Y$ equal 0.80, 0.70, and 0.60. Their covariances with $Y$ are equal to 0.20. The means of the tests are 16.0, 16.0 and 20.0 respectively. Are these tests parallel tests, tau-equivalent, essential tau-equivalent or congeneric? Discuss your answer.

4.8 A test has a mean score equal to 40.0, a standard deviation equal to 10.0 and a reliability equal to 0.5. Which difference score do you expect after a retest when the first score of a person equals 30?

4.9 Two tests $X$ and $Y$ are available. The tests have equal observed-score variances: $\sigma^2_X = \sigma^2_Y = 25.0$. The reliability of test $X$ is 0.8, the reliability of test $Y$ is 0.6. Their intercorrelation is zero. Compute the reliability of the composite test $X + Y$. Also, compute the reliability of the composite after doubling the test length.

# 5 GENERALIZABILITY THEORY

## 5.1 Introduction

An observed score obtained with a psychological or other behavioral measurement instrument is just one of many possible scores that could have been obtained, since an alternative context, other measurement conditions and varying circumstances may lead to other observed scores. In other words, an observed score is usually obtained for a particular test form. Another, equivalent test form, however, may have been as appropriate for our measurement purpose, but might have led to a different observed test score. Consequently, if one wants to model observed scores, one has to take into account many sources of variation (including error variation). This also applies if one considers the reliability of scores obtained from a measurement instrument. Classical reliability provides a decomposition of the observed score into a true score and only one type of error. Theoretically this error is undifferentiated. The several reliability estimation procedures lead to specific conceptualizations of error. Parallel test forms reliability, e.g., considers the lack of equivalence between the forms as the source of error, test-retest reliability the time of testing, and internal consistency reliability the variability in test items.

Generalizability theory or G theory (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach et al. 1972), in contrast to classical test theory, provides a decomposition of an observed score taking into account more sources of variation, dependent upon the specific measurement situation. So G theory also recognizes that multiple sources of error may operate in a measurement implying that there is no unitary definition of reliability. This is, basically, a consequence of the view that a specific test score or any other particular behavioral measurement for that matter (e.g. a job performance score, or an expert's performance assessment of student achievement) is conceived of as a sample from a universe of admissible or suitable observations. Such a universe is characterized by one or more sources of variation, the so-called facets. In Section 5.2 an overview of the basic concepts of G theory will be given. Clearly, in view of how a particular behavioral measurement is conceived, the designs used to collect the measurements come in focus. In section 5.3 and 5.4 some of the more simple one- and two-facet designs will be given, together with the corresponding decomposition of the observed score into a component for the true score, viz. the universe score, and one or more error components. We shall see that with each of the components variances, or rather variance components are associated. An estimate of each variance component can be obtained from an analysis of variance (ANOVA). In section 5.5 an extensive example will be described from a study by Webb, Shavelson, Kim and Chen (1989) on the reliability of job performance measurements. In Sections 5.6 and 5.7 two-facet nested designs are introduced. In Section 5.8 designs with fixed facets are discussed, and in Section 5.9 kinds of measurement errors. In Section 5.10 attention is paid to conditional errors of measurement. Finally, in Section 5.11 some concluding remarks are made.

## 5.2 Basic Concepts of G Theory

A particular behavioral measurement is conceived of as a sample from a *universe* of admissible observations, or a domain of suitable or appropriate observations. The universe or domain is characterized by one or more sources of error variation, called *facets*. In a study where students' performance is rated by judges on performance criteria, judges and performance criteria are the facets of a measurement, and each facet consists of a set of conditions. Usually the facets are assumed to be indefinitely large. The universe, then, is defined as all possible conditions of the facets. Ideally, a person's *universe score* is his or her average score over all conditions. In a measurement situation, however, error is at stake, and this calls for the estimation of variance components. These estimates of variance components are highly informative, whereas the so-called *generalizability coefficients* - the G theory counterparts of classical reliability coefficients - are straightforward ratios of appropriate universe-score variances to total-score variances. In addition to variance components standard errors are considered as appropriate indicators of uncertainty of, e.g., performance assessments of student achievement or school effectiveness (Cronbach, Linn, Brennan & Haertel, 1997).

The purpose of G theory is to generalize from an observation at hand, i.e. the observed score, to the appropriate universe of observations. This domain or universe is defined by all possible conditions of the facets of the study. It should be noticed that the *object of measurement* (e.g., the persons to whom a test is presented) is not a facet.

The measurements in a study are obtained according to a particular design. A one-facet example of a measurement design is an *n*-item test presented to a number of persons. In this example we have a crossed design with all combinations of items ($i$) and persons ($p$); this crossed design is denoted as $p \times i$. If different sets of items are presented to different persons, we have a nested $i : p$ (read $i$ within $p$) design instead. With two facets several (partially) nested designs are possible. Let us take a study with items ($i$) and raters or judges ($j$). In this study all persons ($p$) have answered all items. This part of the study is a crossed $p \times i$ design. The items have been distributed among the judges. In other words, each judge had another set of items to rate: items are nested within judges. Judges have been crossed with persons: if an item has been allocated to a judge, he or she has rated the answers of all persons to this item. This partially nested design is denoted as $p \times (i : j)$. A representation of a nested $p \times (i{:}j)$ design and that of a crossed $p \times i \times j$ design with the same number of measurements for each person is given in Figure 5.1

| | judge 1 | | judge 2 | |
|---|---|---|---|---|
| | item 1 | item 2 | item 1 | item 2 |
| person 1 | × | × | × | × |
| person 2 | × | × | × | × |

$p \times i \times j$

| | judge 1 | | judge 2 | |
|---|---|---|---|---|
| | item 1 | item 2 | item 3 | item 4 |
| person 1 | × | × | × | × |
| person 2 | × | × | × | × |

$p \times (i : j)$

**FIGURE 5.1.** A crossed $p \times i \times j$ design and a nested $p \times (i : j)$ design with the same number of measurements for each person

In G theory two types of facets are distinguished: *random* facets and *fixed* facets. In a random facet the number of conditions is thought to be infinite. That is, the conditions of a facet that are selected for a particular measurement procedure or test are assumed to be a random sample from a very large number of possible conditions. And we would like to generalize over all admissible or suitable observations from the universe. Over fixed facets no generalization is sought, as the number of conditions in a fixed facet is equal to the number of conditions in the study or measurement procedure. A combination of a random facet and a fixed facet leads to a mixed-facet generalizability study.

In G theory a distinction is made between a generalizability or G study, and a decision or D study. A G study investigates the influence of the sampling of conditions from various facets on observed scores. So the purpose of a G study is to collect as much information on the sources of variation as possible. A D study provides information for substantive decisions. It decides on the specific design that is most suitable for typical applications. Clearly, such a decision depends upon several factors. First, the relevant universe of generalization must be defined. Secondly, it must be decided whether the purpose of the study involves so-called relative or absolute decisions. Relative decisions refer to the comparison of a person's achievement, e.g., relative to other persons' achievements. Absolute decisions are made when one is interested in a person's universe score per se. With relative decisions go relative errors, and absolute errors are associated with absolute decisions. A third factor upon which the choice of a D study depends is the size of the various sources of variation. In addition, practical considerations such as the availability of judges, e.g., in a rating study, and costs associated with gathering the data are also criteria for deciding on a specific measurement design. As in a D study alternative possibilities can be tried out with respect to the number of conditions of the facets involved, a D study can be regarded as a generalization of the Spearman-Brown formula for test length.

## 5.3 One-Facet Designs, the $p \times i$ Design and the $i : p$ Design

*The Crossed Design*

Let us have a universe with one facet, the facet *items*. We assume that the facet is a random facet. For person $p$ the observed score on item $i$ is $X_{pi}$. When we want to generalize over the facet, we must take the expectation of $X_{pi}$ over items. This expectation defines the universe score:

$$\mu_p \equiv \mathrm{E}_i X_{pi}. \tag{5.1}$$

The universe score is comparable to the true score in classical test theory. In generalizability theory it is assumed that the persons are a random sample from a large population (formally: $N_p = \infty$). Analogously to definition (5.1) we can define the population mean of item $i$ as:

$$\mu_i \equiv \mathrm{E}_p X_{pi}. \tag{5.2}$$

The expectation of the universe scores is $\mu$. This universe mean also is the expectation of the population means $\mu_i$.

With these definitions we can decompose the observed score $X_{pi}$ into a number of components:

$$\begin{array}{lll} X_{pi} = & \mu & \text{grand mean} \\ & + \mu_p - \mu & \text{person effect} \\ & + \mu_i - \mu & \text{item effect} \\ & + X_{pi} - \mu_p - \mu_i + \mu & \text{residual} \end{array} \tag{5.3}$$

So, the observed score can be written as the sum of the grand mean, a person effect, an item effect and a residual. The residual can be thought of as a combination of pure measurement error and the interaction between the item and the person. But, for lack of replications these two sources of variation are confounded.

In Figure 5.2 a Venn diagram representation is given of the $p \times i$ design. An advantage of such a representation is that it visualizes the variance components involved in the decomposition of the observed scores. The interaction can be found in the segment where the circles for persons and items overlap.

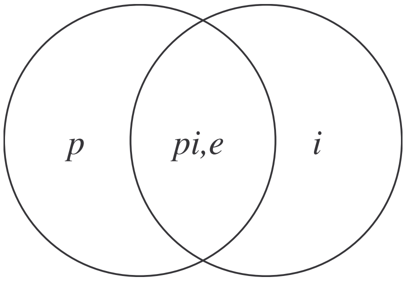


**FIGURE 5.2.** Venn diagram representation of the $p \times i$ design

The population variance of universe scores or person effects is called the variance component for persons and written as $\sigma_p^2$. We have also a variance component for items, $\sigma_i^2$, and a residual variance component, $\sigma_{pi,e}^2$. The notation of the residual reflects the confounding of the random error and the interaction. The variance of $X_{pi}$ over $p$ and $i$, $\mathrm{E}_{p,i}(X_{pi} - \mu)^2$ is

$$\sigma^2(X_{pi}) = \sigma_p^2 + \sigma_i^2 + \sigma_{pi,e}^2 . \tag{5.4}$$

The three variance components can be estimated from an ANOVA of a two-way design. It should be noticed that in the ANOVA terminology two ways are distinguished, one representing the units of measurements, i.e. the persons in G theory, and the other the facet, items. So, a two-way ANOVA design is equivalent to or rather leads to a one-facet G study. The observations in a crossed design can be written down as in Figure 5.3. In the rightmost column we have the averages for persons, averaging over items. In the bottom row we have the average scores for items, averaging over persons.

$$\begin{array}{cccccc}
x_{11} & \cdots & x_{1i} & \cdots & x_{1n_i} & x_{1.} \\
\cdot & & \cdot & \cdot & \cdot & \cdot \\
x_{p1} & \cdots & x_{pi} & \cdots & x_{pn_i} & x_{p.} \\
\cdot & & \cdot & \cdot & \cdot & \cdot \\
x_{n_p1} & \cdots & x_{n_pi} & \cdots & x_{n_pn_i} & x_{n_p.} \\
x_{.1} & \cdots & x_{.i} & & x_{.n_i} & x_{..}
\end{array}$$

**FIGURE 5.3.** Observations in a crossed $p \times i$ design with $n_p$ persons and $n_i$ items

In order to compute the variance components for this $p \times i$ design we use the ANOVA machinery. We start with calculating the sums of squares for persons and items, respectively. For the computation of the sum of squares for persons we replace each entry $x_{pi}$ in the row for a person by the average score for this person. Next we take the squared deviations from the grand mean and sum these squared deviations. The sum of squares for persons is

$$\mathrm{SS}_p = \sum_{p=1}^{n_p} n_i (x_{p.} - x_{..})^2 . \tag{5.5}$$

The mean squares for persons is obtained by dividing the sum of squares for persons by the degrees of freedom corresponding to this sum of squares, $n_p - 1$. The mean squares for persons is equal to $n_i$ times the variance of the mean scores and equal to the total-score variance divided by $n_i$.

The sum of squares for items is obtained in a similar way. The simplest way to obtain the sum of squares for the residual is to compute the total sum of squares and to subtract the sum of squares for persons and the sum of squares for items. The complete ANOVA is summarized in Table 5.1. The rightmost column in this table gives the expected mean squares for the random-effects model. In the expected mean squares for persons all variance components related to persons are included as well as the random error. This is due to the fact that the model is a random model. In a model with fixed effects the interactions for a particular person would have summed to 0. In the model with random effects the $n_i$ interactions are a random sample from all possible interactions for the person. The coefficient of the variance component for persons is $n_i$. This coefficient is equal to the number of observations in which the person effect is involved.

**TABLE 5.1.** ANOVA of the crossed $p \times i$ design

| Source of variation | Sum of squares SS | Degrees of freedom *df* | Mean squares MS | Expected MS: EMS |
|---|---|---|---|---|
| Persons ($p$) | $\mathrm{SS}_p$ | $n_p$ - 1 | $\mathrm{MS}_p\mathrm{=SS}_p\mathrm{/df}_p$ | $\sigma^2_{pi,e} + n_i\sigma^2_p$ |
| Items ($i$) | $\mathrm{SS}_i$ | $n_i$ - 1 | $\mathrm{MS}_i\mathrm{=SS}_i\mathrm{/df}_i$ | $\sigma^2_{pi,e} + n_p\sigma^2_i$ |
| Residual | $\mathrm{SS}_{pi,e}$ | ($n_p$ - 1)($n_i$ - 1) | $\mathrm{MS}_{pi,e}\mathrm{=SS}_{pi,e}\mathrm{/df}_{pi,e}$ | $\sigma^2_{pi,e}$ |
| Total | $\Sigma\Sigma(x_{pi}\text{-}x_{..})^2$ | | | |

$\mathrm{MS}_{pi,e}$ estimates $\sigma^2_{pi,e}$. From the above table we obtain an estimate of $\sigma^2{}_p$:

$$\hat{\sigma}^2_p = (\mathrm{MS}_p - \mathrm{MS}_{pi,e}) / n_i\,. \tag{5.6}$$

The generalizability coefficient for the $n_i$-item test – universe-score variance divided by observed variance - is

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pi,e} / n_i}\,, \tag{5.7}$$

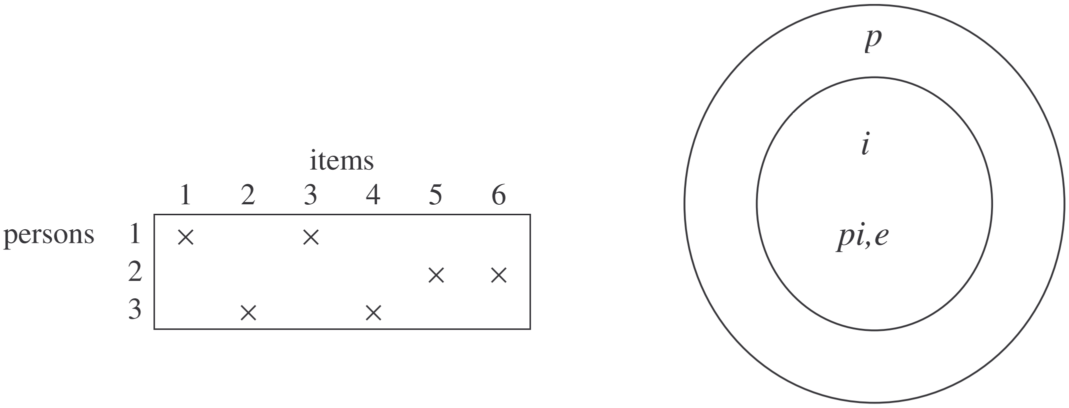
also known as the stepped-up intraclass correlation. The expectation sign indicates that in this formula an approximation is given to the expected squared correlation of observed scores and universe scores. The coefficient is the generalizability counterpart of the reliability coefficient. Its size gives information on the accuracy with which comparisons between persons can be made. So, the coefficient concerns relative measurements, and this is denoted by *Rel* (Shavelson & Webb, 1991). The estimate of (5.7) in terms of mean squares is

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{\mathrm{MS}_p - \mathrm{MS}_{pi,e}}{\mathrm{MS}_p}\,. \tag{5.8}$$

The mean squares in (5.8) can be written in terms of the total-score variance and the item variances. If we do so, we can derive that (5.8) is identical to coefficient α, the coefficient known as a lower bound to reliability. This implies that in case of congeneric items generalizability theory underestimates generalizability or reliability. The problem is due to the fact that the true-scale differences between congeneric measurements are taken up into the interaction term in score composition (5.3).

*The Nested i : p Design*

In the one-facet $i : p$ design each person is presented with a different set of items. This situation is schematized in Figure 5.4. It is clear from the figure that the data matrix is incomplete.



a) Data matrix $i : p$ b) Venn diagram

**FIGURE 5.4.** Data matrix and Venn diagram for the nested $i : p$ design

Only two variance components can be estimated. The ANOVA for the nested design is given in Tabel5.2.

**TABLE 5.2.** ANOVA of the nested $i : p$ design

| Source of variation | Sum of squares SS | Degrees of freedom *df* | Mean squares MS | Expected MS: EMS |
|---|---|---|---|---|
| Persons ($p$) | $SS_p$ | $n_p$ - 1 | $MS_p=SS_p/df_p$ | $\sigma^2_{i,pi,e} + n_i \sigma^2_p$ |
| Residual | $SS_{i,pi,e}$ | $n_p(n_i - 1)$ | $MS_{i,pi,e}=SS_{i,pi,e}/df_{i,pi,e}$ | $\sigma^2_{i,pi,e}$ |
| Total | $\Sigma\Sigma(x_{pi}-x..)^2$ | | | |

With $n_i$ items the generalizability coefficient is

$$\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{i,pi,e}/n_i}, \qquad (5.9)$$

which is estimated by

$$\rho^2_{\text{Rel}} = \frac{\text{MS}_p - \text{MS}_{i,pi,e}}{\text{MS}_p}. \qquad (5.10)$$

Notice the difference between the left hand sides of (5.7) and (5.9). In the nested design the ratio of variance components equals the generalizability coefficient. For more on this design see Jackson (1973).

## 5.4 The Two-Facet Crossed $p \times i \times j$ Design

Now two facets define our universe. Consider a universe with items and judges as facets. We have $n_p$ persons, $n_i$ items and $n_j$ judges. All persons answer all items. Each judge rates the answer of each person to each item. All combinations of $n_p$ persons, $n_i$ items and $n_j$ judges occur. We have a fully crossed $p \times i \times j$ design. The Venn diagram for this design appears in Figure 5.5. The ANOVA for this design with random effects is given in Table 5.3. Now we have a residual consisting of random error and the three-way interaction. There are three two-way interactions and three main effects. Again the coefficients for the variance components in the expected mean squares are equal to the number of times an effect is present in the study. For each person $n_i n_j$ observations are available; therefore the coefficient for the variance component for persons equals $n_i n_j$.

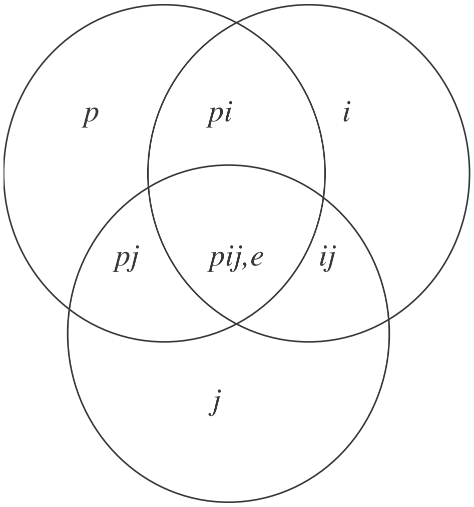


**FIGURE 5.5.** Venn diagram for the two-facet crossed $p \times i \times j$ design

**TABLE 5.3.** ANOVA of the crossed $p \times i \times j$ design

| Source of variation | *df* | EMS |
|---|---|---|
| Persons (*p*) | $n_p - 1$ | $\sigma^2_{pij,e} + n_i\sigma^2_{pj} + n_j\sigma^2_{pi} + n_i n_j\sigma^2_p$ |
| Items (*i*) | $n_i - 1$ | $\sigma^2_{pij,e} + n_p\sigma^2_{ij} + n_j\sigma^2_{pi} + n_p n_j\sigma^2_i$ |
| Judges (*j*) | $n_j - 1$ | $\sigma^2_{pij,e} + n_p\sigma^2_{ij} + n_i\sigma^2_{pj} + n_p n_i\sigma^2_j$ |
| Interaction *pi* | $(n_p - 1)(n_i - 1)$ | $\sigma^2_{pij,e} + n_j\sigma^2_{pi}$ |
| Interaction *pj* | $(n_p - 1)(n_j - 1)$ | $\sigma^2_{pij,e} + n_i\sigma^2_{pj}$ |
| Interaction *ij* | $(n_i - 1)(n_j - 1)$ | $\sigma^2_{pij,e} + n_p\sigma^2_{ij}$ |
| Residual | $(n_p - 1)(n_i - 1)\,)(n_j - 1)$ | $\sigma^2_{pij,e}$ |

Analogously to Formula (5.7) we obtain a generalizability coefficient:

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pi}/n_i + \sigma^2_{pj}/n_j + \sigma^2_{pij,e}/n_i n_j}\,. \tag{5.11}$$

When we set the expected mean squares in Table 5.3 equal to the observed mean squares, we can solve the seven equations for the seven variance components. We start in the bottom row of the table: the residual component is set equal to the observed residual mean squares. Next, we compute the variance component for the interaction between items and judges, etc. It is possible that while doing so negative estimates of variance components are obtained. The best one can do is to compute all components using possibly negative values. After all components have been computed, we set negative values equal to zero (Brennan, 1992). This is the way estimation proceeds in the simultaneous estimation procedure of some software packages. The generalizability estimate in terms of mean squares is

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{\mathrm{MS}_p - \mathrm{MS}_{pi} - \mathrm{MS}_{pj} + \mathrm{MS}_{pij,e}}{\mathrm{MS}_p}\,. \tag{5.12}$$

Formula (5.12) too can be rewritten in terms of variances and covariances. Let us first examine the structure of the variance-covariance matrix in the crossed design. We do this with the help of Figure 5.6.

| | | judge 1 | | judge 2 | |
|---|---|---|---|---|---|
| | | item 1 | item 2 | item 1 | item 2 |
| judge 1 | item 1 | $\sigma^2_{1(1)}$ | $\sigma_{1(1)2(1)}$ | $\sigma_{1(1)1(2)}$ | $\sigma_{1(1)2(2)}$ |
| | item 2 | $\sigma_{2(1)1(1)}$ | $\sigma^2_{2(1)}$ | $\sigma_{2(1)1(2)}$ | $\sigma_{2(1)2(2)}$ |
| judge 2 | item 1 | $\sigma_{1(2)1(1)}$ | $\sigma_{1(2)2(1)}$ | $\sigma^2_{1(2)}$ | $\sigma_{1(2)2(2)}$ |
| | item 2 | $\sigma_{2(2)1(1)}$ | $\sigma_{2(2)2(1)}$ | $\sigma_{2(2)1(2)}$ | $\sigma^2_{2(2)}$ |

**FIGURE 5.6.** Variances and covariances in the crossed design with 2 judges and 2 items

Only the covariances $\sigma_{1(1)2(2)}$ and $\sigma_{1(2)2(1)}$ are 'pure' covariances: covariances between different items, rated by different judges. If we denote pure covariances by $\sigma_{i(j)i'(j')}$, then we can write the following reliability coefficient (cf. Equation 4.10)

$$\rho_{XX'} = \frac{n_i^2 n_j^2 \overline{\sigma_{i(j)i'(j')}}}{\sigma_X^2} = \frac{n_i n_j}{(n_i - 1)(n_j - 1)} \left( 1 - \frac{\sum_{i=1}^{n_i} \sigma_i^2 + \sum_{j=1}^{n_j} \sigma_j^2 - \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \sigma_{i(j)}^2}{\sigma_X^2} \right), \tag{5.13}$$

where $\sigma_i^2$ is the total score variance for item $i$ and $\sigma_j^2$ is the total score variance for judge $j$. Coefficient (5.13) is identical to generalizability coefficient (5.12).

In generalizability theory the emphasis is not so much on the estimation of reliability or, better, generalizability, as on the estimation of the variance components. The relative size of a component gives us information on the influence that this component has on measurement error. After the components have been estimated we can do a D-study. We are then able to compute the generalizability coefficient for a number of items and/or a number of judges different from that in the G-study. We also are able to estimate the effect of using another design to obtain observations. Let us restrict ourselves to the application of the crossed design. Let $n_i'$ be an arbitrary number of items and $n_j'$ an arbitrary number of judges, then the following formula gives the generalizability coefficient which will be obtained for these numbers of items and judges:

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pi}/n'_i + \sigma^2_{pj}/n'_j + \sigma^2_{pij,e}/n'_i n'_j}\,. \tag{5.14}$$

This formula is a generalization of the Spearman-Brown formula for a lengthened test. With the formula we can investigate the effect on generalizability of increasing the number of items and the effect of increasing the number of judges.

Generalizability theory is not the only possibility for reliability estimation with more than one facet. A crossed design might also be analyzed with structural equation modeling; for examples, see Blok (1985), and Werts, Breland, Grandy, & Rock (1980).

## 5.5 An Example of a Two-Facet Crossed $p \times i \times j$ Design: The Generalizability of Job Performance Measurements

Webb et al. (1989) studied the reliability of scores of job performance of Navy machinist mates in the perspective of G theory. Three raters (supervisor, peer, and self) rated the machinist mates on four measures of job performance: a hands-on performance test, a paper-and-pencil job knowledge test, job task performance ratings, and global ratings. G theory was utilized to estimate the reliability of the measures (G study), and to improve the measurement design (D study).

Let us look at one part of the study, the ratings on the hands-on performance tests. Two examiners or raters observed each machinist on eleven tasks in the engine room. Details of the procedure are given by Webb et al. (1989, p. 97-8).

Table 5.4 gives some results of a G study and a D study of the hands-on performance tests in terms of estimated variance components and the generalizability coefficients for relative errors. From the estimated variance components we see that *examiner* was a negligible source of variation. This also holds for the interaction effects of persons and examiners, and tasks and examiners. So it may be concluded that examiners rank machinist mates highly similarly on the hands-on performance tests. The main effect for tasks, however, is relatively high. This means that tasks differ in difficulty. The variance component for the interaction between persons and tasks also is relatively large. It accounts for 60 percent of the variability of the scores $X_{pij}$. So, differences between persons were greater for some tasks than for others. This all has important implications for further improving the measurement of job performance using a variety of tasks (see Webb et al., 1989, p. 100). Most influential for reliability or generalizability is to introduce more tasks. From the table it can be seen what the influence on the G coefficient would be with 17 tasks given only one examiner. Also the generalizability for absolute decisions, a subject that will be discussed in another section, would be largely improved with an increase in the number of tasks.

How would a classical reliability study on the data be carried out? Researchers may differ on which reliability coefficient to calculate. Is a reliability coefficient as such informative in the study described in the example? The answer is no. Estimated variance components are the preferred statistics to compute (cf. *Standards*, APA, AERA, & NCME, 1999, Chapter 2). Not only can everybody then calculate his or her own reliability or generalizability coefficient, there also are clear indications how to improve measurement by varying the “size of the study design”.

**TABLE 5.4.** Estimated variance components and generalizability coefficients for hands-on performance tests (G study and D-study)

| Source of variation | Variance components | | | |
|---|---|---|---|---|
| Persons | .00626 | | | |
| Examiners | .00000 | | | |
| Tasks | .00970 | | | |
| Persons × Examiners | .00000 | | | |
| Persons × Tasks | .02584 | | | |
| Examiners × Tasks | .00003 | | | |
| Residual | .00146 | | | |
| Size of design | | | | |
| Number of examiners | 1 | 1 | 2 | 1 |
| Number of tasks | 1 | 11 | 11 | 17 |
| G coefficient (relative) | .19 | .72 | .72 | .80 |

## 5.6 The Two-Facet Nested $p \times (i : j)$ Design

In Section 5.3 the crossed $p \times i$ design and the nested $i : p$ design have been discussed. A counterpart of the crossed $p \times i \times j$ design is the partially nested design $p \times (i{:}j)$. The crossed and the partially nested design have been presented in Figure 5.1, with the same number of observations for each person. In the crossed design each judge rated all answers to all items. In the nested design each judge had another set of items at his or her disposal to judge the persons.

Apparently, the nested design is less informative than the crossed design. In the crossed design judges can be compared: they judge the same items answered by the same persons. This comparison is not possible in the nested design. The nested design has, however, an unexpected advantage. Each person answers more items in this design. This suggests that the design might result in a higher reliability and, consequently, might be more efficient. We will demonstrate that this is the case. First, we will analyze the nested $p \times (i : j)$ design .

Let us have $n_j$ judges. Each judge rates another set of $n_i$ items; items are *randomly* allocated to judges (When each judge is expert on one subject area and rates only answers to questions pertaining to that particular subject area, the universe itself can be regarded as nested and the variance components are interpreted in a slightly different way). Each person answers $n_i n_j$ items.

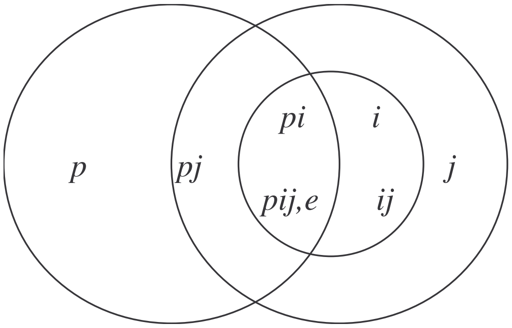


**FIGURE 5.7.** The Venn diagram for the $p \times (i : j)$ design

We can use the Venn diagram in Figure 5.7 as a means of finding which variance components are confounded. Because $p$ and $j$ are crossed, their circles intersect in the diagram. The circles for *p* and *i* intersect too. The circle for *i* lies entirely within the circle for *j*, visualizing the nesting of *i* within *j*. In the diagram *i* and *ij* are found in the same segment, from which we may conclude that these effects are confounded. Also *pi* is confounded with the residual *pij,e*. The ANOVA of the nested $p \times (i : j)$ design is given in Table 5.5.

**TABLE 5.5.** ANOVA of the nested $p \times (i : j)$ design

| Source of variation | *df* | EMS |
|---|---|---|
| Persons ($p$) | $n_p - 1$ | $\sigma^2_{pi,pij,e} + n_i \sigma^2_{pj} + n_i n_j \sigma^2_p$ |
| Judges ($j$) | $n_j - 1$ | $\sigma^2_{pi,pij,e} + n_p \sigma^2_{i,ij} + n_i \sigma^2_{pj} + n_p n_i \sigma^2_j$ |
| Interaction $pj$ | $(n_p - 1)(n_j - 1)$ | $\sigma^2_{pi,pij,e} + n_i \sigma^2_{pj}$ |
| Items within Judges ($i : j$) | $n_j(n_i - 1)$ | $\sigma^2_{pi,pij,e} + n_p \sigma^2_{i,ij}$ |
| Residual | $(n_p - 1)n_j(n_i - 1)$ | $\sigma^2_{pi,pij,e}$ |

We write the generalizability coefficient as

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pj} / n_j + \sigma^2_{pi,pij,e} / n_i n_j} . \tag{5.15}$$

The generalizability coefficient for the nested $p \times (i : j)$ design with $n_i$ items per judge and $n_j$ judges also can be obtained from the results of an analysis with the crossed design. The generalizability coefficient for the nested design, written in terms of the variance components from a crossed analysis, equals

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pj} / n_j + \sigma^2_{pi} / n_i n_j + \sigma^2_{pij,e} / n_i n_j} . \tag{5.16}$$

The contribution of the variance component for the interaction persons × items to the observed variance for persons is smaller than in the crossed design. In the crossed design only $n_i$ items are involved, in the nested design $n_i n_j$. Consequently, the denominator in (5.16) is smaller than the denominator in (5.11) for the crossed design with the same number of observations, and the generalizability coefficient for the nested design is higher than that for the crossed design.

We will not discuss the ANOVA computations for this and other nested designs in further detail. Several statistical software packages enable us to estimate the variance components for other designs than the crossed design.

## 5.7 Other Two-Facet Designs

Four other types of nested designs with two facets can be distinguished:

- $i \times (j : p)$
- $j : (i \times p)$
- $(i \times j) : p$
- $j : i : p$.

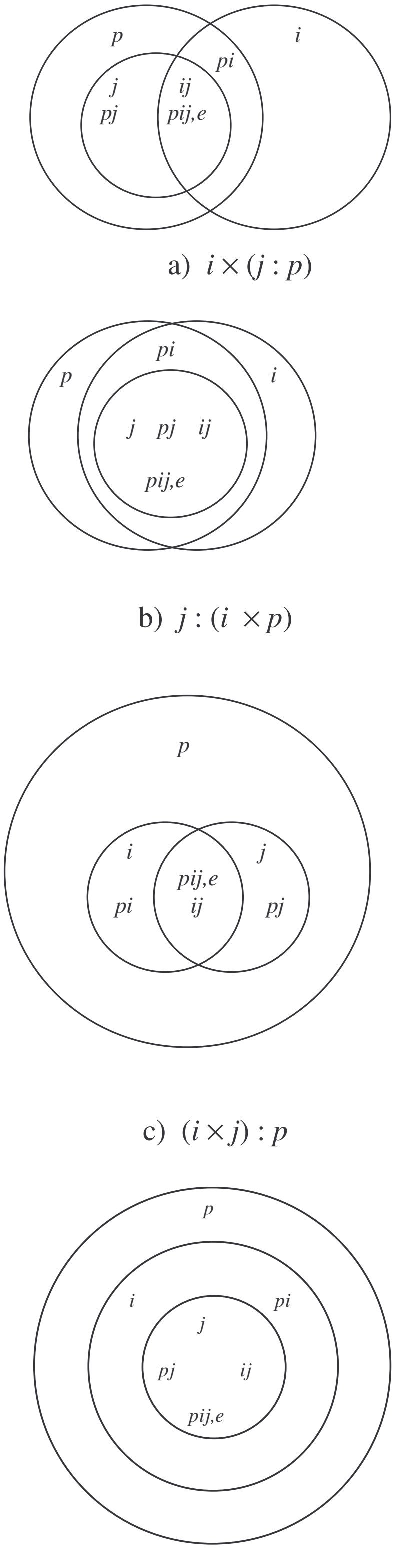
The $i \times (j : p)$ design is formally identical to the $p \times (i : j)$ design. In the $i \times (j : p)$ design the persons and one of the two facets have changed places. So, also in the $i \times (j : p)$ design five variance components can be estimated. An example of the $i \times (j : p)$ design is the design in which the responses to a set of items is judged by a group of judges, and the group of judges differs from person to person.

The $j : (i \times p)$ design and the $(i \times j) : p$ design are not equal formally. These two designs are similar in that in both designs four variance components can be estimated. Let us consider the $o : (i \times p)$ design, where 'o' designates *occasions*. An example of the $o : (i \times p)$ design is a design where each person responds to the same tasks, but the measurement occasions differ, for persons as well as for tasks. In the $(o \times j) : p$ design a group of judges rates the performance of a person, and each person is rated at different occasions and by a different group of judges. For example, person 1 is judged at occasions 1, 2 and 3, by judges 1 and 2, person 2 is judged at occasions 4, 5 and 6, by judges 3 and 4, and so on.

Finally, an example of a fully nested $i : j : p$ design is the situation where each person's work is judged by a different group of judges, and each judge uses another set of tasks $i$. In the $i : j : p$ design only three variance components can be estimated, due to the confounding of many effects.

Cronbach et al. (1972) mention another type of design: a design in which the nested effect $j$ in the combination $(j : i)$ occurs only once, i.e. $n_j = 1$. In this case the notation $(i,j)$ for '$i$ joint with $j$' is used. There are two two-facet designs with '$i$ joint with $j$': the $(i,j) \times p$ design and the $(i,j) : p$ design. In the $(i,j) \times p$ design three variance components can be estimated, in the $(i,j) : p$ design only two. If facet $j$ is considered to have no influence on the score variation, the designs can be simplified: the $(i,j) \times p$ becomes the one-facet $i \times p$ design, and the $(i,j) : p$ design becomes the one-facet $i : p$ design.

The Venn diagrams for the four (partially) nested two-facet designs are presented in Figure 5.8.



a) $i \times (j : p)$

b) $j : (i \times p)$

c) $(i \times j) : p$

d) $j : i : p$

**FIGURE 5.8.** Venn diagrams for four nested designs

## 5.8 Fixed Facets

The definition of measurement error and universe score depends on the extent to which one is willing to generalize over measurement conditions. Let us take the crossed design with facets items and judges. Perhaps we are not interested in generalizing over the judges in the study to a hypothetical universe of judges. Maybe the judges in the study are the only judges available for a long period of time. If we do not generalize over judges we can redefine the effects in such a way that the interactions with judges total to zero. For example, the sum of the person × judge interactions for a particular person equals zero. An alternative way of saying this is that the average person × judge interaction for a particular person is taken up into the universe score for that person. The variance component for the person × judge interaction also becomes part of the universe score variance. The generalizability coefficient for a fixed facet *judges* can be written in terms of the initial variance components as

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{[\sigma^2_p + \sigma^2_{pj}/n_j]}{[\sigma^2_p + \sigma^2_{pj}/n_j] + ([\sigma^2_{pi} + \sigma^2_{pij}/n_j] + \sigma^2_e/n_j)/n_i} . \tag{5.17}$$

Generalizability is higher when a facet is fixed: the numerator in (5.17) is larger than the numerator in (5.11), while the denominator remains the same. This is understandable: generalization over a more limited universe is easier.

The estimated generalizability for the crossed $p \times i \times j$ design with judges fixed equals

$$\mathrm{E}\rho^2_{\mathrm{Rel}} = \frac{\mathrm{MS}_p - \mathrm{MS}_{pi}}{\mathrm{MS}_p} , \tag{5.18}$$

which not only looks quite similar to the corresponding coefficient in the crossed $p \times i$ design, but which in fact gives the same outcome as (5.8). So, if judges in the crossed $p \times i \times j$ design are fixed, we can neglect the facet *judges*: we total over judges and analyze the resulting crossed $p \times i$ design. Why should we analyze the observations as a crossed $p \times i \times j$ design at all? An argument in this particular case to analyze the observations for the full model is that without much difficulty more information is obtained with respect to the relevant variance components.

Also in a nested design facets can be fixed. Let us take an example. A test has been constructed with a number of subtests. The subtests might be tests of different aspects of the subject matter, or various scales that can be distinguished. In a test on test theory subtests might be *classical test theory* and *generalizability theory*. Multiple forms of the test can be constructed, but all tests are to be constructed with the same division into subtests. The facet 'Subtests' is fixed; no generalization is sought over subtests. Items are nested within subtests and persons are crossed with items: the design is a nested $p \times (i : s)$ design. Table 5.5 has been repeated as Table 5.6. The mean squares in Table 5.6 are still specified according to the fully random design. There is a difference between both tables. In Table 5.6 the notation *i*:*s* is used, instead of *i*,*is*. this is done in order to indicate that the nesting is not a result of a design decision, but the result of a nesting of the universe itself: items belong to a specific subtest.

**TABLE 5.6.** ANOVA of the nested $p \times (i : s)$ design

| Source of variation | EMS |
|---|---|
| Persons ($p$) | $\sigma^2_{pi:s,e} + n_i \sigma^2_{ps} + n_i n_s \sigma^2_p$ |
| Subtests ($s$) | $\sigma^2_{pi:s,e} + n_p \sigma^2_{i:s} + n_i \sigma^2_{ps} + n_p n_i \sigma^2_s$ |
| Interaction $ps$ | $\sigma^2_{pi:s,e} + n_i \sigma^2_{ps}$ |
| Items within Subtests ($i : s$) | $\sigma^2_{pi:s,e} + n_p \sigma^2_{i:s}$ |
| Residual | $\sigma^2_{pi:s,e}$ |

When *Subtests* is a fixed facet the variance components must be redefined. The variance components for the model with fixed facet *Subtests* are

$$\sigma^2_{p*} = \sigma^2_p + \frac{\sigma^2_{ps}}{n_s}, \tag{5.19}$$

where $n_s$ is the number of subtests,

$$\sigma^2_{i*} = \sigma^2_{i:s}, \tag{5.20}$$

and

$$\sigma^2_{pi,e*} = \sigma^2_{pi:s,e}. \tag{5.21}$$

One might use the results in Table 5.6 in order to obtain a generalizability coefficient. Unfortunately, the resulting formulas are not very useful when a different number of items is allowed for each subtest or stratum. The generalizability coefficient for a test with fixed subtests or, rather, strata has been derived by Rajaratnam et al. (1965). They proposed an alternative estimate of the generalizability coefficient that is valid also when the number of items varies from stratum to stratum. In the derivation of the formula they used the total-score variance and a weighted sum of the residuals $\text{MS}_{pi,e(s)}$ for the various strata $s$ ($s = 1,\ldots,n_s$). The generalizability coefficient they derived, is

$$\rho^2_{\text{Rel}} = \frac{\sigma^2_X - \sum_{s=1}^{n_s} n_{i(s)} \text{MS}_{pi,e(s)}}{\sigma^2_X}, \tag{5.22}$$

where $n_{i(s)}$ is the number of items in stratum $s$. This coefficient can be rewritten as coefficient $\alpha_s$, the coefficient that already has been introduced in Chapter 4 (4.11).

In the analysis above the scores on the subtests are averaged. Shavelson & Webb (1991) argue that also separate analyses for the subtests should be done. If the results differ strongly between subtests, it might be profitable to use scores on the subtests instead of or in addition to a total score. When scores on the subtests are obtained, we can apply multivariate generalizability theory in order to obtain estimated universe scores defined on the subtests. Cronbach et al. (1972) mention the application of the multivariate approach in connection with the estimation of universe scores for the Wechsler Performance and Verbal scales. Their results are the generalizability theory equivalent of Equation (4.13).

There are two other designs in which a fixed effect makes sense: the $i \times (j : p)$ design and the $j : (i \times p)$ design. In both designs $i$ can be fixed. In both designs the variance component for the interaction between $i$ and $p$ is considered to be part of the universe score variance. It makes no sense to consider a facet as fixed when this facet is nested within a random facet. Fixing both effects in a two-facet design means that one does not want to generalize at all; only pure measurement error is considered to contribute to the error variance in the generalizability coefficient. Generalizability cannot be estimated when both effects are fixed, because pure error variance and at least one interaction variance are confounded.

## 5.9 Kinds of Measurement Errors

Until this moment the variance components for the main effects of the facets have not played a role in the estimation of generalizability for the crossed design. We have argued that in a crossed $p \times i$ design all persons answer the same items; so, the differences between items do not play a role when persons from the same crossed study are compared to each other. In comparing the persons from the same study we are interested in relative measurement.

The size of the item effects is relevant when we have a nested study in which each person responds to a different set of items. The size of the item effects also is relevant if we want to compare the persons in a crossed study with other persons who have taken a different test, or if we want to compare all persons to a standard of performance. In those situations the variation in difficulty level is an interfering factor, to be regarded as part of the measurement error: we have absolute measurement errors.

So far we used the variance for the relative measurement:

$$\sigma^2_{\text{Rel}} = \sigma^2_{pi,e} / n_i \, . \tag{5.23}$$

With absolute errors we have an error variance equal to

$$\sigma^2_{\text{Abs}} = \sigma^2_i / n_i + \sigma^2_{pi,e} / n_i \, . \tag{5.24}$$

Instead of generalizability coefficient (5.7) for relative measurement errors we can define a coefficient for absolute measurement errors as:

$$\varphi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_i / n_i + \sigma^2_{pi,e} / n_i} \, . \tag{5.25}$$

The coefficient is estimated by

$$\hat{\varphi} = \frac{\text{MS}_p - \text{MS}_{pi,e}}{\text{MS}_p + (\text{MS}_i - \text{MS}_{pi,e}) / n_p} \tag{5.26}$$

or by the biased estimator

$$\hat{\phi} = \frac{n_i}{n_i - 1} \left( \frac{s_x^2 - \sum_{i=1}^{n_i} s_i^2}{s_x^2 + n_i^2 s_c^2 / (n_i - 1)} \right) \tag{5.27}$$

where $s_x^2$ is the total score variance (with denominator $n_p$), $s_i^2$ the variance of condition $i$ (with denominator $n_p$) and $s_c^2$ the variance of the condition means (with denominator $n_i$) (Rajaratnam, 1960).

We denote coefficient φ the *index of dependability* following the suggestion by Brennan & Kane (1977). This coefficient cannot be regarded as a correlation except in the nested design. In the nested $i : p$ design relative errors and absolute errors are identical. For a comparison of the designs and coefficients, see Exhibit 5.1.

**Exhibit 5.1. The One-Facet Crossed $p \times i$ Design and the Nested $i : p$ Design**

*design* $p \times i$

estimated variance components:

$$\hat{\sigma}_p^2, \hat{\sigma}_i^2, \hat{\sigma}_{pi,e}^2$$

generalizability coefficient given $n_i$ items (relative decisions):

$$\mathrm{E}\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{pi,e}^2 / n_i}$$

index of dependability given $n_i$ items (absolute decisions):

$$\phi = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_i^2 / n_i + \hat{\sigma}_{pi,e}^2 / n_i}$$

*design* $i : p$

estimated variance components:

$$\hat{\sigma}_p^2, \hat{\sigma}_{i,pi,e}^2 (= \text{estimate of } \sigma_i^2 + \sigma_{pi,e}^2)$$

generalizability coefficient given $n_i$ items

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \hat{\sigma}_{i,pi,e}^2 / n_i}$$

in this nested design there is no difference between absolute and relative decisions

The index of dependability for the crossed design estimates the generalizability for the nested $i : p$ design

The index of dependability has a lower value than the coefficient for relative measurement errors because more variance components contribute to the absolute error. The variance component for items in the crossed $p \times i$ design is based on the assumption of random sampling of items. If items vary in difficulty level we probably are not prepared to sample items randomly from the universe. We will stratify the universe and use a stratified random sampling scheme instead. Within strata items will vary less in difficulty. When items are sampled from a stratified universe the absolute error variance is overestimated with (5.24).

The relative and absolute errors might be viewed as two extreme possibilities to think about when discussing errors and decision making. When using absolute errors we implicitly assume that the scores are not corrected for measurement bias. We might, for example, use a test known to be relatively difficult without correcting the test scores. The alternative is, of course, to use score corrections wherever possible. In many situations enough knowledge with respect to the relative difficulty of alternative tests is available in order to correct scores at least partially.

Let us take the following model:

$$\begin{array}{lll} X_{pi} = & \mu & \text{grand mean} \\ & + \mu_p - \mu & \text{person effect} \\ & + \mu_i - \mu & \text{condition effect} \\ & + X_{pi} - \mu_p - \mu_i + \mu & \text{residual} \end{array} \tag{5.28}$$

The score $x_{pi}$ obtained under condition $i$ can be corrected for the condition effect $\mu_i - \mu$, giving

$$x'_{pi} = x_{pi} - (\mu_i - \mu). \tag{5.29}$$

Suppose that two tests $X$ and $Y$ have been administered to two large random samples from the population. Then the condition means $\mu_i$, the population means of the two tests, are known and scores on both tests can be corrected by taking deviation scores $x_{pi} - \mu_i$. So, test scores are made comparable with a relative measurement approach - an approach that makes use of the results obtained by a group of examinees on a test. In Chapter 10 the conditions under which scores on different tests can be made equivalent, are discussed more extensively.

The correction in (5.29) is an ideal, however. In practice the effects $\mu_i - \mu$ needed for the correction are imperfectly estimated. If the number of persons tested under condition $i$ is very small, we have little information on the size of the value $\mu_i$. Then it is sensible not to correct and the absolute error is the relevant error for absolute decision making. With a larger number of persons the condition mean $x_{.i}$ contains useful information on the value of $\mu_i$. But the influence of measurement error may still be so large that we should not estimate $\mu_i$ by the mean condition score. However, if some information on the variation of condition effects is available, one might estimate the condition effect $\mu_i$ using a Kelley-formula. Such a procedure was suggested by De Gruijter & Van der Kamp (1991), based on work by Lindley (1971) on what is to be regarded as a very simple multilevel model (Snijders & Bosker, 1999). The equation for the estimation of condition mean $\mu_i$ based on the results of $n_i$ persons is

$$\hat{\mu}_i = \hat{\rho}_i^2 x_{.i} + (1 - \hat{\rho}_i^2)\hat{\mu}\,, \tag{5.30}$$

where $x_{.i}$ is the observed mean for condition $i$,

$$\hat{\mu} = \frac{\sum_{j=1}^{m} w_j x_{.j}}{\sum_{j=1}^{m} w_j}\,, \tag{5.31}$$

$$\hat{\rho}_i^2 = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_{p,res}^2 / n_i} \tag{5.32}$$

and

$$w_i = \frac{1}{\hat{\sigma}_b^2 + \hat{\sigma}_{p,res}^2 / n_i}\,. \tag{5.33}$$

An estimate of the within-condition variance is

$$\hat{\sigma}_{p,res}^2 = \frac{\sum_{j=1}^{m}\sum_{p=1}^{n_j}(x_{pj} - x_{.j})^2}{\sum_{j=1}^{m} n_j - m}\,. \tag{5.34}$$

The variance component for conditions can be estimated as

$$\hat{\sigma}_b^2 = \frac{\sum_{j=1}^{m}(x_{.j} - \bar{x})^2}{m-1} - \frac{(\sum_{j=1}^{m}\frac{1}{n_j})}{m}\hat{\sigma}_{p,res}^2\,, \tag{5.35}$$

with

$$\bar{x} = \frac{\sum_{j=1}^{m} x_{.j}}{m} \tag{5.36}$$

(Jackson, 1973).

The formulas allow for conditions with different numbers of persons. The weight of an observed condition mean for the computation of the estimate of the universe mean is relatively high when the number of persons who have been measured under that condition, is large. From (5.30) and (5.32) we see that in this case also the observed mean score for the condition is close to the true condition mean.

Now scores can be corrected for the condition effect. But, for an optimal estimate of a person's universe score Kelley's formula can be used. So, an estimate of the universe score is

$$\hat{\mu}_p = \rho_{\text{Rel}}^2 (x_{pi} - \hat{\mu}_i) + \hat{\mu}\,, \tag{5.37}$$

where $\rho^2{}_{\text{Rel}}$ is the reliability of condition *i*. The error associated with the procedure obviously differs from the absolute error. It also differs from the relative error unless the condition mean and universe mean are perfectly estimated.

The estimation formula can be rewritten as

$$\hat{\mu}_p = \hat{\mu} + \rho_{\text{Rel}}^2 (x_{pi} - x_{.i}) + \alpha(x_{.i} - \hat{\mu})\,, \tag{5.38}$$

where $\alpha$ is the product of the generalizability coefficient and 1 - $\rho_i^2$. Jarjoura (1983, see also Longford, 1994) discusses the optimal estimation of universe scores on basis of a *n*-item test without taking the intermediate step of estimating the condition mean $\mu_i$. When items are randomly selected for test forms, (5.38) is identical to his formula (39).

Related to the sampling approach in generalizability theory is *matrix sampling* (for an overview, see Lord & Novick, 1968, and Sirotnik & Wellington, 1977). In matrix sampling from a population of persons and a one-facetted universe the restrictions in the choice among designs used in generalizability theory can be dropped. In large-scale testing programs, program evaluation and the measurement of group performance matrix sampling is applied. Here the term matrix sampling refers to a measurement format in which a large set of test items is organized into relatively short item sets, each of which is randomly assigned to a subsample of test takers, so avoiding to administer all items to all examinees (*Standards*, APA, AERA, & NCME, 1999, p. 178).

Other designs than the crossed design and the nested design allow us to efficiently estimate condition effects. Figure 5.9 shows a design where judges overlap. So, judge 1 and judge 2 have group 2 in common, but only judge 1 rates the persons from group 1 and only judge 2 rates persons in group 3. Through the overlap all judges are connected and the relative leniency of judges can be estimated. With a design like the design in Figure 5.9 we have already left the domain of generalizability theory. See, for an overview of methods to estimate judgmental effects the contribution by Braun (1988).

| | group 1 | group 2 | group 3 | group 4 |
|---|---|---|---|---|
| judge 1 | × | × | | |
| judge 2 | | × | × | |
| judge 3 | | | × | × |
| judge 4 | × | | | × |

**FIGURE 5.9.** A judgmental design with overlap between judges

## 5.10 Conditional error variance

Another important issue in summarizing reliability data and errors of measurement, not yet discussed in this chapter, is the reporting of conditional error variance (see *Standards*, APA, AERA, & NCME, 1999, p. 27). Generalizability theory allows for a conditional standard error of measurement and a conditional error variance, i.e. the error variance varies over particular levels of scores. It is more likely that the error variance is not constant than that it is constant. Brennan (1998) gives information on the estimation of the conditional error variance, for absolute as well as for relative measurements.

Let us consider the single-facet design. The estimate of the absolute error variance for person $p$ is

$$\hat{\sigma}^2_{\text{Abs}(p)} = \frac{\sum_{i=1}^{n_i}(x_{pi} - x_{p.})^2}{n_i(n_i - 1)}. \tag{5.39}$$

This estimate of the conditional error variance is valid for the crossed $p \times i$ design as well as the nested $i : p$ design. When the items are scored dichotomously, the relevant model is the binomial model and (5.39) can be simplified to (6.3) divided by the square or the number of items.

The estimation of the conditional relative error variance for the $p \times i$ design by Brennan is based on work by Jarjoura (1986). The conditional relative error variance, the variance of $\delta_p$ can be written as

$$\begin{aligned}
&\sigma^2_{\mathrm{Rel}(p)} = \mathrm{var}[(X_{pI} - \mu_I) - (\mu_p - \mu) \mid p] \\
&= \sigma^2_{\mathrm{Abs}(p)} + \sigma^2_i / n_i - 2\,\mathrm{cov}(X_{pI} - \mu_p, \mu_I - \mu \mid p) \\
&= \sigma^2_{\mathrm{Abs}(p)} - \sigma^2_i / n_i - 2\,\mathrm{cov}(\mu_i - \mu, X_{pi} - \mu_p - \mu_i + \mu \mid p) / n_i ,
\end{aligned} \tag{5.40}$$

where $X_{pI}$ and $\mu_I$ are means over $n_i$-item sets. Averaged over persons the latter covariance term is zero and the relation between absolute errors and relative errors as given in (5.23) and (5.24) is obtained. The conditional relative error variance is estimated as

$$\hat{\sigma}^2_{\mathrm{Rel}(p)} = \frac{n_p + 1}{n_p - 1} \hat{\sigma}^2_{\mathrm{Abs}(p)} + \hat{\sigma}^2_i / n_i - 2 \left( \frac{n_p}{n_p - 1} \right) \frac{\mathrm{cov}(x_{pi}, x_{.i} \mid p)}{n_i} . \tag{5.41}$$

Brennan considers other designs as well. One of these designs is the design in which a table of specifications is used for the stratification of items. The estimated conditional absolute error variance for the stratified design is a generalization of the conditional error variance derived by Feldt (1984) for dichotomous items and mentioned in the next chapter.

## 5.11 Concluding Remarks

After more than a century of educational and psychological testing a broadening of measurement, and at the same time a development in sophistication, is visible. In addition to traditional measurement so-called assessments become more important (cf. the *Standards*, APA, AERA, & NCME, 1999). A broadening of applications in practice goes hand in hand with the development of more sophisticated statistical models for test scores obtained in all kinds of assessments and also in program evaluation. G theory can be used in the analysis of sources of variation of assessments. Examples of the use of G theory in the assessment of student achievement and school effectiveness have been given by Cronbach et al. (1997). Looking back at more than four decades of G theory, however, one must come to the conclusion that G theory is underutilized, in spite of recent work in the field.

Recent applications and developments in specific theoretical areas are the analyses of quantitative behavioral observational data (Suen & Ary, 1989; see also Rogosa & Ghandour, 1991). G theory in the context of repeated measurements is only incidentally mentioned (Shavelson, Webb, & Rowley, 1989). Every now and then authors are propagating G theory for designing, assessing and improving the dependability of measurement procedures (e.g., Marcoulides, 1999): hardly to any avail. There are rays of hope for the future, however. The relevance of the use of G theory and G coefficients have found ample place in the 1999 *Standards for educational and psychological testing* (APA, AERA, & NCME, 1999, Chapter 2).

## Exercises

5.1 We have the following table:

| | item |
|---|---|
| person | 1 2 3 4 5 6 7 8 9 10 |
| 1 | 1 1 0 1 1 0 0 1 1 0 |
| 2 | 1 0 1 0 1 1 0 0 0 0 |
| 3 | 0 0 0 1 1 1 0 1 0 1 |
| 4 | 1 1 0 1 0 0 0 0 0 1 |
| 5 | 1 1 1 1 0 1 0 0 0 0 |
| 6 | 1 1 0 0 0 0 0 1 0 1 |
| 7 | 1 0 1 0 1 0 0 0 1 1 |
| 8 | 0 0 0 0 1 1 0 0 0 0 |
| 9 | 1 1 1 1 1 0 1 1 1 0 |
| 10 | 1 1 1 1 1 1 1 1 0 0 |
| 11 | 1 1 1 1 1 0 0 1 0 1 |
| 12 | 1 1 1 0 0 1 1 1 1 1 |
| 13 | 1 1 1 1 1 0 1 1 0 0 |
| 14 | 0 1 1 1 0 0 1 0 1 1 |
| 15 | 1 0 1 1 1 1 1 1 1 0 |
| 16 | 1 0 0 1 0 1 1 0 1 0 |
| 17 | 1 1 0 1 1 0 1 1 1 0 |
| 18 | 1 1 1 1 1 1 0 1 1 1 |
| 19 | 1 1 1 1 1 1 1 1 0 1 |
| 20 | 1 1 1 1 1 1 1 1 0 1 |

Compute the item variances and the variance of total scores. Next, compute coefficient α. Compute the mean squares for items, persons and interaction. Compute the variance components and discuss the implications of the values of these components. Finally, compute the generalizability coefficient. Use statistical software if you want to.

5.2 A test consisting of fifteen open answer items is given to 500 examinees. The responses are judged by four judges in a completely crossed design. The mean squares from an ANOVA are given in the table below. Compute the variance components and the generalizability coefficient for 15 items and 4 judges.

| Source of variation | MS | Source of variation | MS |
|---|---|---|---|
| Persons (*p*) | 17.30 | *pj* | 0.80 |
| Items (*i*) | 1051.65 | *ij* | 45.65 |
| Judges (*j*) | 420.80 | *pij,e* | 0.65 |
| *pi* | 6.65 | | |

5.3 Compute the generalizability coefficient for a) 30 items and 4 judges, and b) 60 items and 2 judges, using the estimated variance components from exercise 5.2.

5.4 The following table gives the expected mean squares for the nested $j : (i \times p)$ design. Give the coefficients of the variance components in terms of $n_p$, $n_i$ and $n_j$.

EMS of the nested $j:(i \times p)$ design

| | |
|---|---|
| $\text{EMS}_p$ | $\sigma^2_{j,pj,ij,pij,e} + a\sigma^2_{pi} + b\sigma^2_p$ |
| $\text{EMS}_i$ | $\sigma^2_{j,pj,ij,pij,e} + c\sigma^2_{pi} + d\sigma^2_i$ |
| $\text{EMS}_{pi}$ | $\sigma^2_{j,pj,ij,pij,e} + e\sigma^2_{pi}$ |
| $\text{EMS}_{j,pj,ij,pij,e}$ | $\sigma^2_{j,pj,ij,pij,e}$ |

5.5 Derive the formula for the correlation between two judges who both judge the responses to $n_i$ items. Use the notation of the variance components from generalizability theory.

5.6 Derive the formulas for the relative and absolute error variance for the crossed $p \times i \times j$ design.

5.7 Three judges have rated 50 examinees each. The variances of the ratings are practically equal for all three judges. The pooled within-judges variance equals 100.0. The judges have different means. Judge 1 has a mean equal to 32.0, judge 2 has a mean equal to 35.0 and judge 3 has a mean equal to 38.0. Is a correction for the difference in leniency indicated? If so, how should we correct the scores?

# 6 MODELS FOR DICHOTOMOUS ITEMS

## 6.1 Introduction

The simplest items in achievement testing e.g. only have two different outcomes, *correct* and *incorrect*. These items are dichotomous. If an examinee does not answer an item, we evaluate the non-response as *incorrect*. A correct answer can be assigned a score 1, and an incorrect answer can be assigned a score 0 (see Exhibit 6.1). Dichotomous items are frequently used in tests. For example, achievement, aptitude and intelligence tests with multiple-choice items are frequently scored dichotomously. For dichotomous and dichotomized items test models have been developed to account for the scores of persons on such tests.

The first model to be discussed – in Section 6.2 - is the binomial model. This model is relevant for the nested *i* : *p* design. It also is the adequate model if items are psychometrically exchangeable. In Section 6.3 the generalized binomial model for items with varying difficulties is introduced. The generalized binomial model is the unidimensional model for dichotomous items within the context of true-score test theory. It is worth to discuss the model, for it spans the bridge to item response models (Chapter 8). Section 6.4 relates the generalized binomial model to the item response models. In Section 6.5 the relevance of item statistics for item analysis and item selection is clarified.

**Exhibit 6.1. On the Existence of Dichotomous Items**

Dichotomous items as such do not exist, dichotomous scoring does. The responses to items do not fall naturally into two categories, "correct" and "incorrect". It takes a decision to code nonresponse and incorrect response(s) all in the same category.

In tests with multiple choice items sometimes a scoring formula is used in order to suppress pure guessing. The possible scores are 1 (correct), 0 (omit, not reached) and -1/(*k*-1), where *k* is the number of response options. When a test is not speeded a "deterrent" against guessing is not likely to be very effective: a person should always respond to an item if he/she has some partial knowledge.

Items also can be weighted. A correct answer on one item might, for example result in a score of 1 point, whereas a correct answer to another item might result in two score-points. Empirical weighting of dichotomous items will be discussed in connection with maximum likelihood estimation of person parameters (Chapter 8). Formula scoring implicitly weights two-choice items heavier than four-choice items, even though two-choice items are less accurate in the lower score-range than four-choice items.

## 6.2 The Binomial Model

If we throw an unbiased dice, the probability of obtaining the outcome five or six equals 1/3. When we throw the dice again, the probability again equals 1/3. The probability of having *x* times the outcome five or six in *n* throws is given by the binomial distribution with parameter $\zeta$ = 1/3:

$$f(x\,|\,\zeta) = \binom{n}{x}\zeta^{x}(1-\zeta)^{n-x}\,, \qquad (6.1)$$

where

$\binom{n}{x} = \frac{n!}{(n-x)!x!}$ is the binomial coefficient with $n! = n(n-1)...1$.

In this section we shall develop the binomial model, assuming the existence of a large item pool. The items are assumed to be independent: the correct answer to one item does not give away the correct response to another item. We randomly select one item from the item pool and ask a person to answer this item. The probability that this person answers a randomly selected item correctly is called his or her domain score $\zeta$. In other words, when we repeat the testing procedure the expected value of the proportion correct answers equals $\zeta$.

Let us administer not one, but $n$ randomly selected items. The probability of a correct answer is equal to $\zeta$ for each of the randomly selected items. The probability of exactly $x$ items correct out of $n$ is given by the binomial model presented in Formula (6.1) (actually the binomial model is an approximation for $n$ smaller than infinity if we use item selection without replacement). With a large number of repeated selections of $n$-item tests the empirical distribution of the number correct will approximate the distribution defined by (6.1).

The result of our exercise is that we have a strong true-score model with respect to the distribution of observed scores (and errors) on the basis of a few weak assumptions. The model is called strong because the error distribution given the domain score is known. There are no assumptions besides the assumption of a large item pool, and the random selection of items from this pool. It is possible for a person to know some of the items from the pool and to answer those items correctly. He or she may not know the correct answer to other items and guess correctly or not when answering these items. It might even be possible for the test administrator to know which items will be answered correctly and which will not be answered correctly. To illustrate this, suppose that a person has to respond to items on addition and subtraction. All addition items are correctly answered and none of the subtraction items. If the next item is presented and this item turns out to be an addition item we assume that the person will answer this item correctly. Nevertheless, whether we have some information or not, over replications of $n$-item tests the distribution of total score will be the binomial distribution.

Now let us consider the situation of a large item pool with more persons. If we give these persons the same selection of $n$ items, it is unlikely that the binomial model holds. From the responses it will become clear that the items have different psychometric characteristics. For one thing, they are likely to differ in difficulty level.

When more persons are tested the binomial model still holds if for every person a separate random selection of items from the item pool drawn. In terms of generalizability theory: we have a nested $i : p$ design.

The binomial model has been popular in educational testing (Hambleton & Novick, 1973). In educational testing frequently a large domain of real or hypothetical items can be constructed and a test can be viewed as a random item selection from this item pool. The purpose of testing is to obtain an estimate of the domain score  (universe score in terms of generalizability theory). Relevant questions are to which extent the person has achieved mastery of the domain, and whether the amount of mastery is enough to pass the person on the examination. In terms of generalizability theory: one is interested in absolute measurement.

An alternative to random selection of items is using a stratified sampling scheme. In relatively heterogeneous item domains we are likely to prefer this sampling scheme. In a relatively homogeneous item domain we might actually be prepared to select items randomly from an item pool. We will elaborate this latter possibility.

*The Binomial Model in a Homogeneous Item Domain*

In the binomial model the variance of measurement errors given test length $n$ and domain score ζ, which is the variance of observed scores given $n$ and ζ, equals

$$\sigma^2_{X|\varsigma} = n\zeta(1-\zeta)\,. \tag{6.2}$$

With an $n$-item test the true score of person $p$ is $\tau_p = n\zeta_p$. However, in this case it is more convenient to keep using the true-proportion correct scale ζ. An application of the binomial model with observed-score variance (6.2) is given in Exhibit 6.2.

The error variance of person $p$ can be estimated from the number correct score $x_p$ as

$$\hat{\sigma}^2_{E(p)} = \hat{\sigma}^2_{X(p)} = n\left[\frac{x_p - n(x_p/n)^2}{n-1}\right] = \frac{x_p(n-x_p)}{n-1}\,. \tag{6.3}$$

The error variance is small for domain scores close to 0 and 1, and high for domain scores close to ½. It is clear that the assumption of an error variance independent of the true score-level is untenable. The estimated conditional standard error of measurement on the proportion correct scale - the square root of (6.3) divided by $n$ - can be used to construct a confidence interval for $\zeta_p$. Due to the fact that the binomial errors are asymmetrically distributed around ζ, and the fact that the size of the variance varies with ζ, the construction of a confidence interval for ζ unfortunately is not straightforward (see Pearson & Hartley, 1970). For not too extreme proportions correct $\bar{x}_p = x_p/n$ and for not too small test sizes $n$ a normal distribution can be used for the computation of a confidence interval around $\bar{x}_p$.

There is a second reason not to trust a confidence interval based on the observed proportion correct blindly. When we are dealing with a population of persons such a confidence interval may well be misleading. We have to take the population distribution into account in the construction of such an interval. For a comparable situation we refer to the discussion around the Kelley-formula in Chapter 3.

What are the characteristics of the procedure with randomly selected $n$-item tests? How to express reliability for the procedure in terms of the ratio of true-score variance and observed-score variance for a particular population of persons? Let us first estimate the average error variance. Using (6.3) we can estimate the error variance related to observed scores $X$ through averaging the estimated error variances for all persons. We obtain

## Exhibit 6.2. Minimum Test Length

Consider the following problem. We have an ability level $\varsigma_h$ that is considered as a definitely high level and another ability level $\varsigma_l$ that is low. We want to classify an examinee as a high ability examinee when $x \geq x_0$ and as low otherwise. We want to have an error probability $P(x < x_0|\zeta_h) \leq \alpha$ for a specified high ability $\zeta_h$. We also want to have an error probability $P(x \geq x_0|\zeta_l) \leq \beta$ for a specified low ability $\varsigma_l$. How many items are needed to achieve the specified accuracy, and for which cut score $x_0$? We will discuss the simpler problem with $\beta = \alpha$.

The minimum test length is the smallest number of items $n$ for which

$$\min_{x_0} \{\max[P(x < x_0 \mid \zeta_h), P(x \geq x_0 \mid \zeta_l)]\} \leq \alpha \, .$$

When $n$ is not too small and the ability $\varsigma$ not too extreme the distribution of $x$ can be approximated by a normal distribution with mean $\zeta$ and standard deviation $n^{½}\sigma_\zeta = n^{½}[\zeta(1\text{-}\zeta)]^{½}$. Let $z_\alpha$ be the $z$-score corresponding to the cumulative probability $\alpha$ in the normal distribution. Then $x_0$ and $n$ can be obtained from the equations

$$\frac{x_0 - n\zeta_h}{\sqrt{n}\sigma_{\zeta_h}} = z_\alpha$$

and

$$\frac{n\zeta_l - x_0}{\sqrt{n}\sigma_{\zeta_l}} = z_\alpha \, .$$

The minimum test length is

$$n = z_\alpha^2 \frac{(\sigma_{\zeta_h} + \sigma_{\zeta_l})^2}{(\zeta_h - \zeta_l)^2}$$

and the corresponding cut score

$$x_0 = n \frac{\sigma_{\zeta_h}\zeta_l + \sigma_{\zeta_l}\zeta_h}{\sigma_{\zeta_h} + \sigma_{\zeta_l}} \, .$$

Birnbaum (1968, pp. 448-449) and Fhanér (1974) give a more general treatment of the subject. Unfortunately the normal approximation does not always give the correct result: the minimum number of items tends to be underestimated. Part of the problem is that $x_0$ in the approximation is a continuous variable. For better results, the cut score should take on an integer value minus a continuity correction equal to ½. Wilcox (1976) demonstrated that an exact solution for the binomial model is feasible.

In Chapter 9 another solution to the problem of minimum test length is discussed, within the framework of IRT.

$$\sigma_E^2 = \frac{1}{N}\sum_{p=1}^{N}\hat{\sigma}_{X(p)}^2 = \frac{1}{n-1}[n^2\mu_{\bar{x}}(1-\mu_{\bar{x}}) - \sigma_X^2\,], \tag{6.4}$$

if – in the computation of the observed-score variance – the numerator is divided by the number of persons $N$ instead of the usual $N$ – 1. In the above formula $\mu_{\bar{x}}$ equals the proportion correct averaged over persons. This results in reliability coefficient:

$$\mathrm{KR}_{21} = \frac{n}{n-1}\left(1 - \frac{n\mu_{\bar{x}}(1-\mu_{\bar{x}})}{\sigma_X^2}\right). \tag{6.5}$$

This coefficient is known as the Kuder-Richardson Formula 21, in a crossed design a lower lower bound to reliability than KR20 (coefficient $\alpha$). Here, in the nested design, the formula does not give a lower bound but is exact, apart from sampling fluctuations.

The Kelley formula for the estimation of the domain score is given by

$$\hat{\zeta}_p = \mathrm{KR}_{21}\bar{x}_p + (1 - \mathrm{KR}_{21})\mu_{\bar{x}}\,. \tag{6.6}$$

The regression of domain scores on observed scores or proportions is linear if the population distribution is given by a beta distribution (see Novick & Jackson, 1974). We have a linear regression with unequal error variances! In Chapter 3 linear regression of true scores on observed scores was obtained for equal error variances. If the domain scores have a beta distribution, we not only have an exact point estimate of $\zeta_p$ (Formula 6.6), but the complete posterior distribution (see Exhibit 6.3).

The nested design in which for each examinee a different random sample of items is selected, is easily implemented on the computer. With computerized testing it is also easy to adapt the test length. If, after the administration of a number of items, the estimate of the domain score is accurate enough, testing can be stopped. A very simple stopping rule has been suggested by Wilcox (1981). Wilcox assumed that there is a test procedure with a fixed test length of $n$ items. An examinee passes the test if at least $n_c$ items are answered correctly. This procedure can be adapted as follows:

- stop after $n_c$ correct responses
- stop after $n$ - $n_c$ +1 incorrect responses.

With this procedure test length can be much shorter than $n$ for most examinees. The flexibility of test length is not the only characteristic of the suggested procedure, however. The procedure also assumes that each presented item has been answered. It is not possible to skip an item temporarily or to change the response to an item. An other adaptive procedure, a procedure with an optimal selection of items instead of random sampling, is discussed in Chapter 9.

## Exhibit 6.3. The Beta-Binomial Complex

The beta distribution for domain scores is defined by

$$f(\zeta) \propto \zeta^{a-1}(1-\zeta)^{b-1}, \text{with } a,b > 0.$$

Let us assume that the population distribution of domain scores is the beta distribution with parameters *a* and *b*. A person from the population answers *x* items from an *n*-item test correctly. The probability of *x* correct out of *n* for a particular value of ς is given by

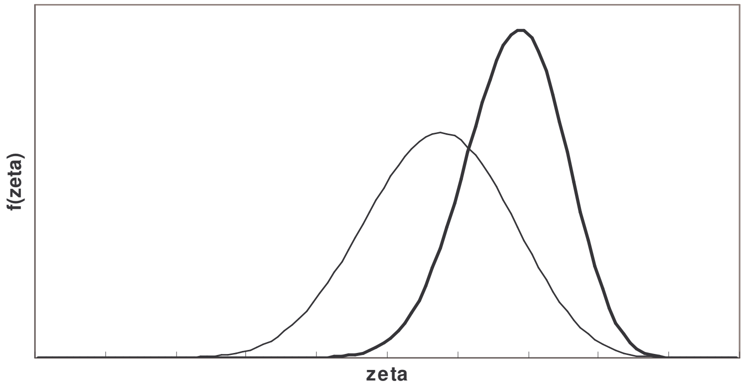
$$f(x \mid \zeta) = \binom{n}{x} \zeta^{x}(1-\zeta)^{n-x} .$$

Notice the similarity of the beta distribution and the binomial distribution. We can derive that the posterior distribution of ζ given the test score is

$$f(\zeta \mid x) \propto \zeta^{a+x-1}(1-\zeta)^{b+n-x-1} ,$$

which is a beta distribution as well. A confidence interval for ζ given the observed score can be obtained; in the literature this kind of confidence interval has been designated a *credibility* interval or a *tolerance* interval.

In the figure the distribution with the larger variation is the beta distribution with *a* = 13 and *b* = 10; its mean equals .57. A person answers 16 out of 20 items correctly; the proportion correct is .8. The more peaked distribution gives the posterior distribution of ς given the score on the 20-item test. Its mean equals .67.



The beta distribution is also used for the construction of the exact "classical" confidence interval for ζ (Pearson & Hartley, 1970).

### *The Binomial Model in a Heterogeneous Item Domain*

In a large heterogeneous item bank the procedure for estimating the domain score, error variance and reliability is as follows. Instead of sampling items randomly from this item bank we randomly select items from various strata. With *q* strata, we randomly select $n_i$ items from stratum *i*. The domain score of interest is then given by

$$\zeta . = \sum_{i=1}^{q} n_i \zeta_i / n , \tag{6.7}$$

with

$$n = \sum_{i=1}^{q} n_i \ ,$$

i.e. the domain score ς. is a weighted average of the domain scores for the various strata (for a more general approach, see Jarjoura & Brennan, 1982). This domain score generally differs from the domain score in (6.1). To illustrate the point, assume that the strata differ in average item difficulty. Also assume that for all strata the same number of items $n_i$ is selected. When the strata sizes are equal, domain score (6.7) equals the domain score under random sampling. The sizes of the strata are arbitrary, however. Some strata might contain more items than other strata, e.g. it might be easier to construct many items for some strata than for other strata. When strata differ in size, the domain score based on stratified sampling can deviate from the domain score under random sampling. Under these circumstances an analysis based on the stratified sampling plan is indicated.

The error variance in the stratified sampling approach equals:

$$\sigma^2_{X|\zeta.} = \sum_{i=1}^{q} n_i \zeta_i (1 - \zeta_i) \, , \tag{6.8}$$

which generally is smaller than the variance that obtains under random sampling. The estimated error variance for person $p$ equals

$$s^2_{E(p)} = \sum_{i=1}^{q} \frac{x_{pi}(n_i - x_{pi})}{n_i - 1} \, . \tag{6.9}$$

(Feldt, 1984). The relevant reliability coefficient is the stratified version of KR21:

$$\mathrm{KR21}_{(\mathrm{s})} = \frac{\sum_{i=1}^{q} \mathrm{KR21}_{(i)} \sigma^2_{Y_i} + \sum_{i=1}^{q} \sum_{j \neq i}^{q} \sigma^2_{Y_i Y_j}}{\sigma^2_X} \, , \tag{6.10}$$

where KR21($i$) is the reliability estimate for the subtest of stratum $i$, and $Y_i$ designates subtest $i$. We should keep in mind here that each subtest contains different items for different examinees.

## 6.3 The Generalized Binomial Model

We start again with an $n$-item test, and this time the $n$-item test is presented to a group of persons. Assuming that the number of items $n$ and the number of persons $N$ are relatively large, we are going to do some computations. We compute the observed proportion correct for an item, say item $i$, within each score group on the test. Next we plot these proportions against the test scores $x$. We may expect to find that the proportion correct responses on the item increases with increasing test score $x$. The result will look like the plot in Figure 6.1. We can do the same thing for a second item, item $j$. It may turn out that items $i$ and $j$ are practically uncorrelated within each score group. We then conclude that the answers to these items are determined by one common factor. This common factor or latent trait score is represented by the true score on the test and is reasonably well approximated by the observed score on the test (If this is the case one actually should expect a slightly negative correlation between the two items in each score group, for the scores on the items must add to *x* in score group *x*; for a nonparametric test for unidimensionality, see Stout, 1987).

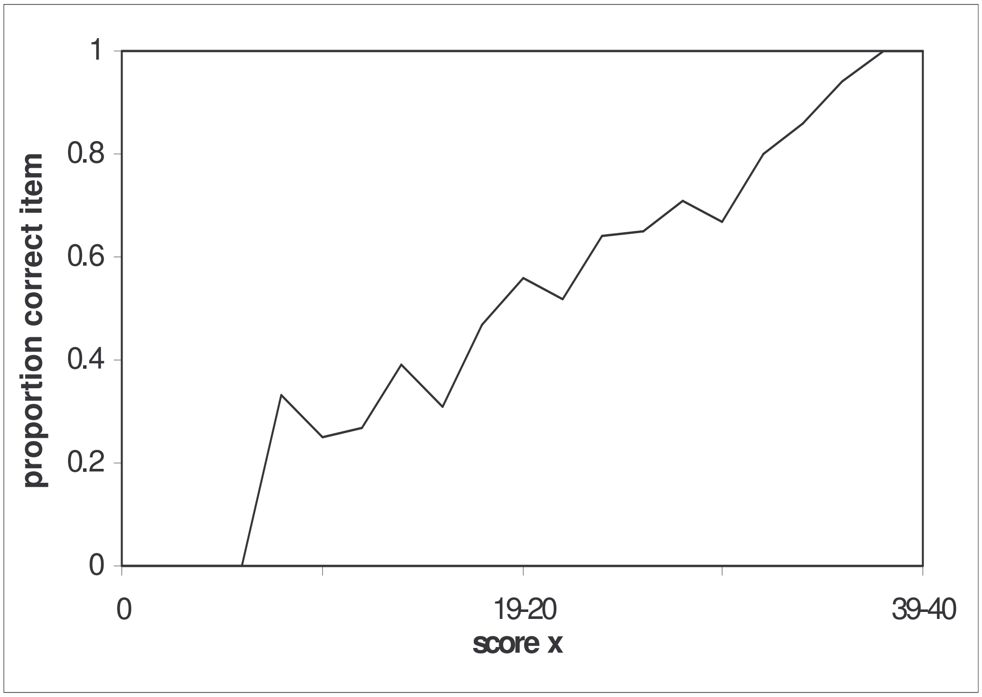


**FIGURE 6.1.** Item-test regression as might be obtained in practice

Again we use the true proportion correct on the test and denote this proportion correct by ζ. The proposition that there is one factor underlying the responses to the test items can be formalized as:

- the probability of a correct answer on item *i* is $P_i(\zeta)$
- the true score on the proportion scale is $\zeta = n^{-1}\Sigma\, P_i(\zeta)$
- Given the true score ζ the responses to items are independent. This is the property of local independence.
  For two items *i* and *j* local independence means:

$$P(X_i = x_i, X_j = x_j \mid \zeta) = P(X_i = x_i \mid \zeta)P(X_j = x_j \mid \zeta)$$

$$= P_i(\zeta)^{x_i}[1-P_i(\zeta)]^{1-x_i} P_j(\zeta)^{x_j}[1-P_j(\zeta)]^{1-x_j}, \qquad (6.11)$$

where $x_i$ equals 1 for a correct answer on item *i* and 0 otherwise. Formula (6.11) is shorthand for: the probability on items *i* and *j* correct is equal to the probability that *i* is correct times the probability that *j* is correct, etc.

Tacitly, but inevitably, we seem to have introduced a strong assumption concerning the process of answering items. From the idea that responses are locally independent it seems to be implied that answering the items is probabilistic. This conclusion is, however, not so inevitable as it appears. Whether or not the answer process is probabilistic can only be verified in a replication study with the same test (cf. the confounding of specific factors and errors in Section 4.4, and the confounding of interaction and error in Chapter 5). For reasons of convenience we will speak of probabilities.

The model that has been introduced above, is the generalized binomial test model (Lord & Novick, 1968). The error variance given ζ defined on the scale of total scores is

$$\sigma^2_{X|\zeta} = \sum_{i=1}^{n} P_i(\zeta)[1-P_i(\zeta)] = n\zeta(1-\zeta) - \sum_{i=1}^{n}[P_i(\zeta)-\zeta]^2 = n\zeta(1-\zeta) - n\sigma^2_{P|\zeta} \qquad (6.12)$$

.

If the item difficulties in the generalized binomial model differ slightly for each level of ζ, the generalized binomial model can be approximated well by the binomial model. This is clear from (6.12). With small differences between items the rightmost factor in (6.12) can be dropped. The more items differ with respect to difficulty, leading to a larger item variance given ζ, the smaller the error variance in the generalized binomial model is relative to the error variance of the binomial model. Does this mean that for accurate testing tests should be used with spread item difficulties? This question is not easy to answer because a different choice of items results in another true score scale. Actually, later in this chapter it is argued that the answer should be 'no' in most cases.

The error variance in the generalized binomial model varies strongly with true score. Can a reasonable estimate of error variance (6.12) be obtained for various levels of ζ? For extreme values of ζ (values close to 0 or 1) the value of (6.12) is close to 0. It seems acceptable to approximate (6.12) by

$$\sigma^2_{X|\zeta} \approx nk\zeta(1-\zeta), \qquad (6.13)$$

with $0 \le k \le 1$. Keats (1957) proposed to choose the factor $k$ so as to be able to reproduce the reliability coefficient $r_{XX'}$ that has been obtained for the test. In this case the estimate of the error variance of person $p$ equals

$$\hat{\sigma}^2_{E(p)} = kx_p(n-x_p)/(n-1), \qquad (6.14)$$

with

$$k = \frac{1-r_{XX'}}{1-\mathrm{KR}_{21}}. \qquad (6.15)$$

Feldt, Steffen & Gupta (1985) compared various methods for the estimation of the variance of measurement errors as a function of true score, including the method proposed by Keats. We will discuss one of the other methods in the next section. Another recent discussion of conditional standard errors of measurement and conditional error variances can be found in Lee et al. (2000).

## 6.4 The Generalized Binomial Model and Item Response Models

The generalized binomial model in (6.11) is a general one-factor model for dichotomous items. The probability of a correct answer to an item increases as a function of true score in a way that is not specified. True score itself is a function of the items and, therefore, is arbitrary. If we would consider to include an item, say item *i*, in a different test, we would have another true score $\zeta'$, monotonously related to the true score of the present test. The function $P_i(\zeta')$ would have another form than the function $P_i(\zeta)$.

The true score scales of the different tests all can be considered functions of one underlying latent trait. Let us denote the latent trait value by the symbol $\theta$. Now we can write the probability of a correct response to item *i* as $P_i(\theta)$. The function $P_i(\theta)$ does not depend on the test form in which item *i* happens to be included. It is assumed that the function $P_i(\theta)$ depends on a number of item parameters. Several one-factor models for dichotomous items have been proposed, like the Rasch-model (Rasch, 1960), and the two-parameter and three-parameter logistic models (Birnbaum, 1968). These models are examples of unidimensional *item response models* (IRT models); there are also multidimensional item response models and models for more than two response categories (some examples will be given in Chapter 8).

The probability of occurrence of a particular response pattern on a *n*-item test given the latent trait score $\theta$ can be written as the product

$$P(X_1 = x_1,\ldots,X_n = x_n \mid \theta) = \prod_{i=1}^{n} P_i(X_i = x_i \mid \theta)$$

$$= \prod_{i=1}^{n} P_i(\theta)^{x_i}[1 - P_i(\theta)]^{1-x_i}\,, \qquad (6.16)$$

where $x_i = 1$ for a correct response and $x_i = 0$ for an incorrect response.

We can estimate the item parameters of the item characteristic curve (ICC), $P_i(\theta)$, of item *i* from responses to the test. Next, we can compute the true score for a given value of $\theta$ as

$$\tau = \sum_{i=1}^{n} P_i(\theta)\,, \qquad (6.17)$$

The conditional error variance for a given true score can be computed as

$$\sigma^2_{E|\tau} = \sigma^2_{E|\theta} = \sum_{i=1}^{n} P_i(\theta)[1 - P_i(\theta)]\,, \qquad (6.18)$$

where $\theta$ is the latent ability that corresponds with the true score.

Further, it is possible to estimate the population distribution of $\theta$ (Bock & Aitkin, 1981). When an estimate of the population distribution is available, we can compute a Bayesian point estimate of $\theta$.

## 6.5 Item Analysis and Item Selection

In traditional item analysis the most common indices that are computed are the indices for item difficulty and item discrimination power. We can do likewise for a nested design as well as for a crossed design. Here, we discuss the computation of item statistics within the context of a crossed design with $N$ persons and $n$ items.

For each item, we can compute the mean item score. For dichotomous items the mean score is equal to the proportion correct, or item difficulty index $p_i$. The higher the value of the item difficulty index, the easier the item is. The variance for item $i$ is

$$Np_i(1-p_i)/(N - 1) \approx p_i(1-p_i). \tag{6.19}$$

The extent to which the item discriminates between high scoring persons and low scoring persons, the item's discriminating power, is approximated by the item-test correlation $r_{it}$. With relatively large tests total test score is close to the true score, and the item-test correlation gives a fair impression of the item discriminating power. With small tests we have a problem. The correlation between item and test, $r_{it}$, is *spurious*: the measurement errors of the item and the test are correlated because the item is part of the total test. In this situation it is better to use $r_{ir}$, the correlation between the item and the rest score, the total score minus the item score. This coefficient can be written as

$$r_{ir} = \frac{s_t r_{it} - s_i}{\sqrt{s_t^2 - 2s_i s_t r_{it} + s_i^2}}\ . \tag{6.20}$$

When in the computation of the variances the numerator is divided by $N$, the item-rest correlation $r_{ir}$ of dichotomous items can be written as

$$r_{ir} = \frac{M_+^{(i)} - M^{(i)}}{\sqrt{s_t^2 - 2s_i s_t r_{it} + s_i^2}} \sqrt{\frac{p_i}{1-p_i}}\ , \tag{6.21}$$

where

$p_i$ = the item proportion correct or item difficulty of item $i$,

$M^{(i)}$ = the average score on the test minus item $i$

$M^{(i)}{}_+$ = the average score on the test minus item $i$ for the subgroup with item $i$ correct.

A coefficient corrected for spuriousness and attenuation has been suggested by Henrysson (1963), with coefficient $\alpha$ as estimator of test reliability.

In a homogeneous test the two item indices, item difficulty and item-rest correlation, give us information on the quality of the item. If necessary, screening of items can be done using these two indices, at least when the sample is large enough to give relatively accurate sample estimates of these indices. In a heterogeneous test, a test from which several subtests can be constructed, the item-rest correlation is less informative. With heterogeneous tests consisting of several subtests factor analysis methodology and possibly structural equation modeling, are approaches that might be useful for test construction and test development in general, and item analysis and item selection in particular. This, however, is beyond the scope of the present chapter (see e.g. McDonald, 1999).

The item-rest correlation $r_{ir}$ should have at least a positive value, the higher the values of the correlation, the better. An item with a value close to 0 may suppress reliability when included in the test, if an unweighted sum score is used. The advantage of unweighted scores is that they are simple, easy to defend and not sensitive to sample fluctuations. Optimal weights might be obtained from an IRT analysis. Of course items with a low discriminating power might be rejected for selection in a final test version.

Although IRT models are discussed in Chapters 8 and 9, here already some remarks will be made about some of the dichotomous IRT models in the context of item analysis and item selection.

In the Rasch model all items are assumed to be equally discriminating. Item selection within the Rasch model involves selecting items with similar item discriminations. In item selection relatively undiscriminating items are deleted from the test, because they do not fit the model. However, also an item with a better than average discrimination will be rejected in a Rasch analysis. Is this desirable from a practical point of view?

What is the optimal difficulty level of test items? Is it good to have items that differ strongly in difficulty level or not? The answer to this question depends on the purpose of the test and the discriminating power of the items. Let us assume that the purpose is to discriminate well in a population of persons. Let us also assume that the items are strongly discriminating. Then the probability of a correct answer shows a large increase at a particular level of the latent trait. In Figure 6.2 we have two such items. The probability of a correct answer on item 1 is close to 1 for levels of the latent trait for which the probability of a correct answer on item 2 is still close to zero. These two items define a Guttman scale as long as no other items of intermediate difficulty are chosen for inclusion in the scale. In the perfect Guttman scale the probability of a correct answer is zero or one: at a particular level of the latent trait the probability jumps from zero to one. That is to say, the Guttman model, leading to the perfect Guttman scale, is a pathological probability model or deterministic model for dichotomous item responses. The Guttman model can also be conceived of as a typical proto-IRT model.

For comparison with Figure 6.2 two less discriminating items are displayed in Figure 6.3.

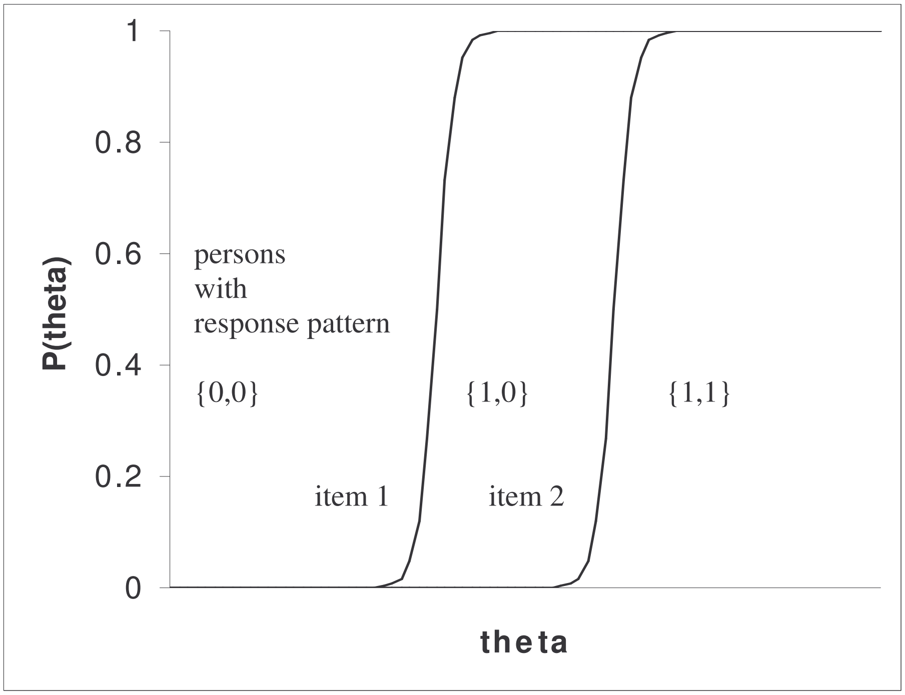


**FIGURE 6.2.** Two strongly discriminating items

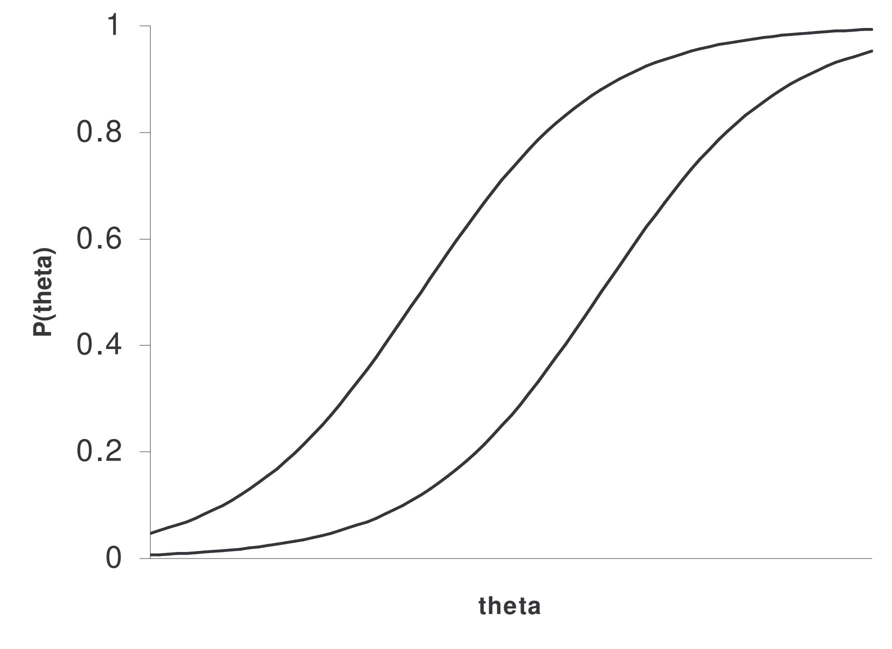


**FIGURE 6.3.** Two items with a moderate discriminating power

If we want to discriminate between persons within a broad range of θ, we better choose items of distinct difficulty levels when we have highly discriminating items like those in Figure 6.2. Each item then contributes to a finer discrimination within a group of persons. The group of persons with 2 items correct out of two can be divided into two subgroups by including a third item that is more difficult than item 2.

In practice most items are more similar in discriminating power to the items in Figure 6.3 than to the items in Figure 6.2. An impression of the discriminating power of items can be obtained by plotting the item-test regression in a figure like Figure 6.1. In case all items would have been Guttman items, the item-test regression would have looked quite differently from the regression in Figure 6.1.

With more moderately discriminating items it proves to be better to select items with comparable difficulties. If we want to discriminate between persons in a population an item difficulty of about 0.50 is optimal unless guessing plays a role (Cronbach & Warrington, 1952). A test with this kind of items is less accurate for persons with very high and very low latent trait values, but for most persons the test is more accurate than a test with spread item difficulties. Some item selection procedures, however, automatically select items with spread difficulties. In a procedure for scale construction proposed by Mokken (1971) the scale is not formed by deleting items that are not satisfactory, but by step-wise adding items that satisfy certain criteria. The procedure starts with the selection of the items most different in difficulty if the items do not differ with respect to discriminating power (see Mokken, Lewis & Sijtsma, 1986); see Croon (Croon, 1991) for an alternative procedure. More information on procedures for test construction is presented in Exhibit 6.4.

### Exhibit 6.4. Item selection in Test Construction: Some Practical Approaches

Traditional test construction relies heavily on the indices for item difficulty and item discriminating power. In addition, item correlations can also be taken into account in the construction of tests. Also, if some external criterion is available, item validity, i.e. the correlation of item and criterion scores can be used.

Several methods have been proposed to construct a relatively homogeneous test from a pool of items. One possible classification of methods is the following:

1. Step-wise elimination of single items or subsets of items. Eliminate those items that do not correlate with the other items (e.g. set a certain threshold for an acceptable item-rest correlation). repeat the procedure after elimination of items until the desired standard is reached. Also the contribution of the item to test reliability can serve as a criterion for elimination of an item.
2. Step-wise addition of single items or subsets of items. The construction of the scale starts with the two items that have the strongest relationship according to a particular index. Next, the item with the strongest relation to the items in the scale in formation is added if certain conditions are satisfied. The process is repeated until no further items are eligible for inclusion. The whole process can be repeated with the construction of the next scale from the remaining items. Another technique in this class of procedures is hierarchical cluster analysis – based on e.g. the average intercorrelation between clusters (see also Nandakumar et al., 1998). In hierarchical cluster analysis scales are constructed simultaneously.
3. Item selection can be based on item correlation with an external criterion. The external criterion can be a classification in a diagnostic category (e.g. schizophrenics). Although this procedure produces a useful instrument for diagnostic purposes, it does not guarantee the construction of a homogeneous scale.
4. Factor analysis of item intercorrelations. Usually this approach is applied when several factors are thought to underlie item responses. Traditional factor analysis is sometimes difficult to apply with dichotomous items. An obvious way out is using one of the procedures of nonlinear factor analysis (Panter, Kimberly & Dahlstrom, 1997). Nonlinear factor analysis can be viewed as a multidimensional IRT analysis. IRT will be outlined in Chapters 8 and 9.

If guessing plays a role, we should take that into account. Let us have a test with multiple-choice items. An item has $k$ response options and one of these is correct. Let us further assume that a person knows the answer to an item and responds correctly or does not know the answer and guesses randomly. The probability of a correct response under random guessing equals $c = 1/k$. Then the relation between $p'$, the item difficulty under guessing, and $p$, the item difficulty without guessing, equals:

$$p' = c + (1 - c)p. \tag{6.22}$$

From this it follows that the optimal difficulty for a multiple-choice item with four options is about 0.625. Actually the optimal value is likely to be somewhat higher (Birnbaum, 1968).

In case we are interested not so much in discriminating between persons as in comparing persons with a standard, the answer to the question of optimal item difficulty is a different one. Assuming that we are interested in a fine discrimination in the neighborhood of a true score level equal to $\tau_0$ an items has an optimal difficulty if:

> at $\tau_0$ the probability of a correct answer equals $c_i + (1\text{-}c_i)\times 0.5$ for $c_i$ equal to 0 and a bit higher if $c_i$ is larger than 0.

This result is obtained with an IRT analysis (Birnbaum, 1968). Such an analysis is to be preferred to an analysis within the context of classical test theory. The true score scale is defined in terms of the items that constitute the test. If one item is dropped from the test, the true score on the test changes as well as the value of $\tau_0$. Later on we will discuss test construction more fully in terms of IRT (see Chapter 9).

The outcome of an item analysis in classical test theory not only depends on the test that includes the item. Also the sample of persons who have answered the test determines the estimates of item discriminating power and item difficulty. It is important to remember this when evaluating test results from different groups of examinees. The groups might differ with respect to performance level and, consequently, an item might have a different estimate of difficulty level in each group. Item selection and test construction on basis of test statistics such as proportion correct is not justified when the estimates for different items come from incomparable groups.

## Exercises

6.1 Compute the probability that a person with a domain score equal to 0.8 answers at least 8 out of 10 items correct, assuming that the items have been randomly selected from a large item pool.

6.2 a. Compute the proportion correct and the item-rest correlation of item 8 in the table of exercise 5.1. Compute the item-test regression of this item.
b. Compute the item-rest correlations of the remaining items as well. Which item should be dropped first when a scale is constructed by a step-wise elimination of items?

6.3 In a testing procedure each examinee responds to a different set of 10 items, randomly selected from a large item pool. The test mean equals 7.5, de standard deviation equals 1.5. What might be concluded about the test reliability?

6.4 For a person $p$ the probability of a correct answer to two items is $P_1(\varsigma_p) = 0.7$ and $P_2(\varsigma_p) = 0.8$, respectively. Compute the probabilities of all possible response patterns.

6.5 What information would you like to obtain in order to verify whether the assumptions made by Keats, see (6.14) and (6.15), are realistic?

6.6 A test consists of three items. The probabilities correct for person $p$ are $P_1(\varsigma_p) = 0.6$, $P_2(\varsigma_p) = 0.7$ and $P_3(\varsigma_p) = 0.8$. Compute the error variance on the total score scale. Also compute the error variance under the binomial model assumption. Comment on the difference.

6.7 Compare $r_{it}$ and $r_{ir}$ for tests with all item variances equal to 0.25 and all interitem covariances equal to 0.05. Compute the correlations for test lengths 10, 20 and 40.

# 7 VALIDITY AND VALIDATION OF TESTS

## 7.1 Introduction

In scientific inquiry validity of statements refers to the degree to which there is empirical evidence to support the adequacy and appropriateness of these statements. More specifically, a measure is valid to the extent that it measures what is intended to measure. How vague this description may be, it makes clear that validity is not a property of a measurement instrument, but rather the interpretation of test scores and their use. In other words, how adequate and appropriate are the interpretations and uses of test scores, taking into account empirical evidence and, eventually, theoretical rationales. Validation, then, is the process through which the validity of the proposed interpretation of scores is investigated. So the process of validation amounts to collecting empirical evidence to provide stable and generally accepted theoretically based interpretations of test scores (and other modes of assessment, for that matter).

In the present chapter we shall first go into the problem of validity as a specific term, i.e. validity of psychological and educational test scores or other assessments. Following recent developments in the conceptualization of validity, validity as a unitary concept will be highlighted. Unified validity integrates earlier forms of validity of test scores by focussing on the various sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes. In Exhibit 7.1 a sketch will be given with respect to validity and the *Standards*, in Section 7.2 the various sources of evidence. How important these sources may be, in the present monograph it is impossible to deal with them extensively. We stick to the statistical aspects of validity and validation. In Section 7.3 selection effects in validation studies are outlined, in Sections 7.4 and 7.5 classification, and in Section 7.6 on what is nowadays called the evidence based approach using analyses of the relationship of test scores to variables external to the test (*Standards*, APA, AERA, & NCME, 1999, pp. 13-4). Some remarks will be made on validation and IRT in Section 7.7, while research validity, a form of validity not typical for test scores, appears in Section 7.8. An important topic that is not included is validity generalization. Suffice it to mention *The Handbook of research synthesis* (1994) edited by H. Cooper & L. V. Hedges, and *Methods of meta-analysis* (1990) by J. E. Hunter & F. L. Schmidt. And it is, of course, a selection from the avalanche of monographs on research synthesis, validity generalization and meta-analysis that have been published at the end of the last century.

### Exhibit 7.1. The Many Faces of Validity & the *Standards*

Where it all started is not easy to trace. Validity, it seems, has been and will be a perennial theme to scrutinize and discuss for test theorists and practitioners in the field of psychological and educational testing. Let us only mention a few highlights in the history of the conceptualization, operationalization and periodic canonization of test validity.

In the fifties of the twentieth century diverse forms of validity have been proposed to fit different situations. The APA (1954) and the AERA (1955) mention four types of validity: content, predictive, concurrent, and construct. Although there was consensus, there were dissidents: Anastasi (1954) added face validity, factorial validity, and various types of empirical validity, Mosier (1947) analyzed face validity into validity by assumption, validity by definition, the appearance of validity, and validity by hypothesis. Interestingly, Ebel (1961) already stressed the evidence base of validity (more explicitly formulated decades later in the 1999 *Standards*). The seminal paper of the fifties, with the benefit of hindsight, is L. J. Cronbach and P. E. Meehl's *Construct validity of psychological tests* (1955).

In 1966 the APA published the Standards for educational and psychological tests and manuals, explicitly using the term standard, i.e. level or degree of quality that is considered proper or acceptable. So, also later editions of the *Standards* provide a frame of reference to assure that relevant issues are addressed. Also in the 1974 Standards (APA, AERA, & NCME, 1974) the distinction in four types of validity mentioned above remains, now also endorsed by AERA and NCME.

A next step in the long march toward a unified view of validity is the 1985 *Standards* (APA, AERA, & NCME, 1985), greatly expanding the formalization of professional standards for test use. No longer are types of validity distinguished, but rather categories of validity evidence called content-related, criterion-related, and construct-related evidence of validity. Messick (1989, pp. 18-20) sketches the historical trends in conceptions of validity, and again, in 1994 made his last public plea for a unified view of validity (Messick, 1995), culminating in the 1999 *Standards* (APA, AERA, & NCME, 1999).

The later codification in 1999 is useful. But we must keep in mind that although a unified view of validity is surely a great stride in the long march, and although a listing of sources of validity evidence (see Section 7.2 and *Standards* 1999 (pp. 11-7) is illuminating and useful, a validation study is always an empirical piece of research according to general rules of research methodology. Apart from the fact that the purpose of the *Standards* is to provide criteria for the evaluation of tests, testing practices, and the effects of test use, it is not a manual for how to set up a validation study with explicit consideration of specific designs and statistical analyses.

## 7.2 Validity and Its Sources of Evidence

"Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests". This is the opening sentence of the chapter on validity in the latest edition of the *Standards* (APA, AERA, & NCME, 1999, p. 9). It is no definition in the Aristotelian sense, i.e. *per genum proximum et differentiam specificam*. Neither is it an operational definition: it does not explicitly refer to the relevant operations to ensure validity. The *Standards* therefore proceed by stating: "The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations" (l.c., p. 9). Fortunately, this is an explicit statement: the unified view of validity entails that validity is evidence based, and the sources of evidence are:

- test content
- response processes
- internal structure
- relations to other variables
- information on the consequences of testing;
  the latter evidence has to do also with social policy and decision making.

The evidence based on test content can be obtained by analyzing the relationship between a test's content and the construct it is intended to measure. Response processes refer to the detailed nature of performance. It generally comes from analyses of individual responses (e.g., do test takers use performance or response strategies; are there deviant responses on certain items, etc.). The evidence based on internal structure comes from the analysis of the internal structure of a test (e.g., can the relationships among test items be accounted for by a single dimension of behavior?). In Chapter 3 we already met the analysis of the internal structure of test items in the context of internal consistency reliability. And the latter form of reliability is worked out (and liberalized, so to say, from the assumptions of classical test theory) in the broader framework of generalizability theory. G theory, therefore, bridges the gap between reliability and validity (cf. Cronbach et al., 1972). Performance assessment is generally thought to have the right content, but performance testing still needs further validation (Messick, 1994).

The largest category of evidence is evidence based on relations to other variables. This category of evidence analyzes the relationship of test scores to external variables (e.g. measures of the same or similar constructs, measures of related and different constructs, performance measures as criteria). Instead of using the old-fashioned label of concurrent validity, e.g., the concept of validity in the unified view refers to the way how evidence can be obtained for validity. The category based on relations to other variables includes:

- convergent and discriminant evidence
- test-criterion relationships
- validity generalization.

The first subcategory of convergent and discriminant evidence has its early beginnings with Cronbach & Meehl (1955), and most importantly, with Campbell & Fiske (1959). This subcategory of what was called construct-related validity is presented in Section 7.5. Test-criterion relationships studies what earlier was called criterion-related validity, and still earlier predictive validity. Validity generalization is the evidence obtained by giving a summing-up of earlier findings with respect to similar research questions (e.g. of the findings of criterion-related correlation studies, with the same or comparable dependent and independent variables). Validity generalization is also known under the terms meta-analysis, research synthesis, or cumulation of studies.

So far, it is all rather abstract. How can it be made more concrete, how do we proceed in the validation of a test? Ironically, to make it clear how we study validity empirically, we do better to go back to the 1985 *Standards* trichotomy of test validity.

The three validities in the 1985 *Standards* are:

1 content-related validity

In general content-related evidence demonstrates the degree to which the sample of items, tasks or questions on a test is representative of some defined universe or domain of content.

2 criterion-related validity

Criterion-related evidence demonstrates that scores are systematically related to one or more outcome criteria.

In this context the criterion is the variable of primary interest, as is determined by a school system, the management of a firm, or clients, for example. The choice of the criterion and the measurement procedures used to obtain criterion scores are of central importance. Logically, the value of the criterion-related study depends on the relevance of the criterion measure that is used.

3 construct-related validity

> The evidence classed in the construct-related category focuses primarily on the test score as a measure of the psychological characteristics of interest. Reasoning ability, spatial visualization, and reading comprehension are constructs, as are personality characteristics such as sociability and introversion. Such characteristics are referred to as constructs because they are theoretical constructions about the nature of human behavior (APA, AERA, & NCME, 1985, pp. 9-11).

Each of these validities leads to methods for obtaining evidence for the specific type of validity. The methods for content-related validity, for example, often rely on expert judgments to assess the relationship between parts of the test and the defined universe. This line of thinking or approach is embedded in generalizability theory as discussed earlier. In addition certain logical and empirical procedures can be used as well (see e.g. Cronbach, 1971).

Methods for expressing the relationship between test scores and criterion measures vary. The general question is always: how accurate can criterion performance be predicted from test scores? Depending on the context a given degree of accuracy is judged high or low, or useful or not useful. Two basic designs can be distinguished for obtaining information concerning the accuracy of test data. One is the *predictive study* where test data are compared, i.e. its relationships are studied with criterion scores obtained in the future. The second type of study is the so-called *concurrent study* in which test data and criterion data are obtained simultaneously.

The value or utility of a predictor test can also be judged in a decision theory framework. This will be exemplified in a later section. There, errors of classification will be considered as evidence for criterion-related validity.

Empirical evidence for the construct interpretation of a test may be obtained from a variety of sources. The most straightforward procedure would be to use the intercorrelations among items to support the assertion that a test measures primarily or substantially a single construct. Technically, quite a number of analytical procedures are available to do so, e.g., factor analysis, MDS, IRT models. Another procedure would be to study substantial relationships of a test with other measures that are purportedly of the same construct and the weaknesses of the relationships to measures that are purportedly of different constructs. These relationships support both the identification of constructs and the distinctions among them. This quite abstract formulation is taken from the *Standards* (APA, AERA, & NCME, 1985, p. 10). In a later section the so-called multitrait-multimethod approach to construct validation will be considered more concretely and in more detail.

Before going into certain aspects and procedures for validation studies, it is important to consider the problem of selection and its effects on the correlation between e.g. test $X$ and criterion $Y$, i.e. the (predictive) validity of test $X$ with respect to criterion $Y$. Essentially, this is applying statistics in the field of psychometrics: what is the influence of restriction of range on the value of the validity of a test?

## 7.3 Selection Effects in Validation Studies

Suppose we want to study the validity of test $X$ with respect to criterion $Y$. We already use the test for selection. Only persons with a score $x$ larger than or equal to score $X_c$ on the test are admitted or selected and we have criterion scores $Y$ only for these selected persons. We are interested in the correlation between $X$ and $Y$ within the total population, but we can compute the correlation only for the subpopulation of selected persons. Is it possible to estimate the correlation within the total population?

We will derive the relation between the correlation in the subpopulation and the correlation in the total population following a practice suggested by Gulliksen (1950). Lowercase characters designate statistics in the subpopulation, capitals designate statistics in the total population. We assume that the regression of *Y* on *X* is linear and that the regressions are identical in the total population and the subpopulation. In addition we assume that the variances of estimation errors, i.e. the variances around the regression line of *Y* on *X*, are identical. These two assumptions can be expressed mathematically as

$$\frac{S_Y}{S_X} R_{XY} = \frac{s_y}{s_x} r_{xy} \tag{7.1}$$

and

$$S_Y^2(1 - R_{XY}^2) = s_y^2(1 - r_{xy}^2)\,. \tag{7.2}$$

Now we have two equations with two unknowns, the criterion variance in the total population $S^2{}_Y$ and the correlation between predictor *X* and criterion *Y* in the total population.

We can solve (7.1) for $S^2{}_Y$ and substitute the result in (7.2). Next, we can solve this equation for $R_{XY}$. The result is

$$R_{XY}^2 = \frac{r_{xy}^2}{r_{xy}^2 + \dfrac{s_x^2}{S_X^2}(1 - r_{xy}^2)}\,. \tag{7.3}$$

Selection on test *X* does not only depress the correlation of this test with the criterion. Due to incidental selection on other variables the correlations of other variables are affected as well. The pattern of correlations differs between the subpopulation and the total population; see Gulliksen (1950) or Mulaik (1972) for the case of more than two variables in selection. The lowering of the correlation of the explicit selection variable *X* sets this test at a disadvantage when it is compared with a competing test in the selected group.

The value of the correlation between two measurement instruments in a subpopulation can be smaller than the correlation in the total population even though no explicit selection has taken place. Self-selection of persons has a similar effect as selection on the extent to which two variables are correlated. Let us give an example of a situation where both selection and self-selection might operate. Assume that mathematical ability is an important ability for a particular study. It is reasonable to assume that there is a relationship between mathematical ability and achievement in the study. We correlate achievement with an ability measure only to find a low correlation. The low value does not invalidate the hypothesis of a relationship. The low correlation might be due to the combined effects of selection and self-selection.

## 7.4 Validity and Classification

The size of the correlation between a predictor and a criterion sometimes says very little about the utility of the predictor (Taylor & Russell, 1939). We will discuss this using Figure 7.1.

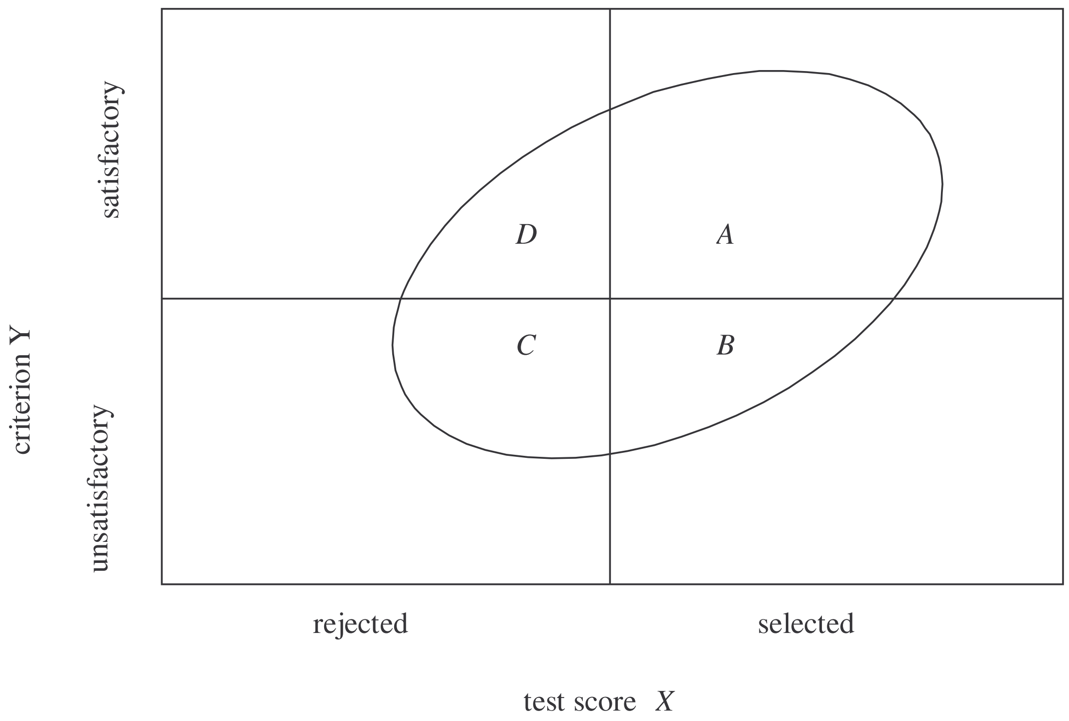


**FIGURE 7.1.** Classification into satisfactory/unsatisfactory and rejected/accepted

Assume that we want to hire a fixed percentage of applicants on basis of their scores on a predictor *X*. For a high score on the predictor we accept the applicant, for a low score we reject the applicant. We have a criterion *Y* with the categories *satisfactory* and *unsatisfactory*.

We have four different outcomes of the selection procedure with corresponding proportions:

- The proportion *A* is accepted and satisfactory
- The proportion *B* is accepted but is unsatisfactory
- The proportion *C* is correctly rejected
- The proportion *D* is rejected although these applicants are satisfactory.

For the sake of simplicity we will not bother about the way in which we can make the distinction between the two groups of rejected applicants. The proportion $A + B$ is called the *selection ratio*; this proportion was assumed to be fixed in the example. The proportion $A + D$ is known as the *base rate*; naturally the base rate is fixed. The proportion $(A+C)$ is called classification accuracy.

Taylor and Russell assumed that a bivariate normal distribution underlies the double dichotomy *satisfactory/unsatisfactory* and *rejected/accepted*. The correlation between the continuous variables *X* and *Y* is denoted by $r_{XY}$. Given a base rate equal to 0.50, a selection ratio equal to 0.30 and a validity coefficient $r_{XY}$ of the underlying continuous variables equal to 0.50 a *success ratio* $R = A/(A + B)$ equal to .74 is obtained. We might compare this outcome with the expected outcome if we did not have used predictor *X*. If we had selected the persons randomly we would have obtained a success ratio equal to the base rate (0.50). We might compare the success ratio also with the success ratio that we would have obtained with a perfect predictor. In the present case the success ratio with a perfect predictor would have been equal to 1.00. The utility of the test as a selection instrument is higher than we might have expected from the size of the validity coefficient.

The results obtained by Taylor and Russell depended very much on their choice of the success ratio as measure of test efficiency and the fact that the selection ratio was fixed. We have two kinds of errors: incorrectly accepting persons and incorrectly rejecting persons. In the Taylor & Russell approach only one kind of classification error counted: incorrectly accepting applicants. In many other applications both kinds of classification errors should be considered and then the utility of the test may be quite different. Let us from now on assume that both kinds of classification errors are relevant and that the selection ratio is not fixed.

Taylor and Russell divided persons from a single population into two groups on basis of criterion performance. Sometimes we are dealing with two or more different populations and we have to decide to which population a person belongs. For this purpose we might use a predictor *X*. In its most simple form the problem comes down to decide on a cut score on the predictor. Persons with a score equal to the cut score or higher are classified as belonging to one of the populations, the others are classified as belonging to the other population. In Figure 7.2 the classification problem is illustrated for the case of two populations.

Let us assume that we have two populations: population I and population II. The purpose of the test is to detect population-II persons or type-II persons, as they need treatment. Let us assume that we have a cut score $X_c$. All persons with a score equal to the cut score or higher than the cut score are classified as type-II persons. We make classification errors. Some persons are incorrectly classified as type-I persons, others are incorrectly classified as type-II persons. With a low cut score we have a relatively large number of false positives, persons who do not belong to population II, but are incorrectly classified as type-II persons. The term ‘positive’ originates from medicine where it indicates the presence of a condition (illness) for wich is screened. In our example ‘positive’ is associated with relatively high test scores. For a high cut score we have, in contrast, a relatively large number of false negatives, persons for whom the diagnosis ‘type-II’ has been missed. At the score *X* at which the curves for the two distributions meet, the two kinds of errors are in balance. We might choose this value of *X* as $X_c$.

The point *X* at which the probability of belonging to population II given the observed score, the posterior probability $P(\mathrm{II}|x)$, equals the posterior probability of belonging to population I depends on the base rate $P(\mathrm{II})$. The way base rate influences the posterior probabilities is easily seen by applying Bayes’ theorem

$$P(\mathrm{II} \mid X) = \frac{P(X \mid \mathrm{II})P(\mathrm{II})}{P(X)} = \frac{P(X \mid \mathrm{II})P(\mathrm{II})}{P(X \mid \mathrm{I})P(\mathrm{I}) + P(X \mid \mathrm{II})P(\mathrm{II})}. \tag{7.4}$$

One might imagine that it is difficult to detect type-II persons when the base rate is low. The accuracy of the measurement instrument with respect to the detection of type-II persons is called its sensitivity. Sensitivity is defined as the proportion of persons with a disorder for whom the diagnosis is correctly made. Not only the sensitivity of the measurement instrument is important, also its specificity. The specificity of a measurement instrument is defined as the proportion of healthy persons for which the diagnosis is correctly rejected. The definitions of the two concepts can be illustrated with the data in Figure 7.3. On the left hand side of this figure the base rate is 0.20. The diagnosis (+) is correctly made in 14 of the 20 cases. So, the sensitivity of the procedure is 0.70. The diagnosis is correctly rejected in 72 of the 80 cases. So, the specificity of the procedure is 0.90. On the right hand side of the figure the base rate is 0.50. The sensitivity is 0.80. The specificity is also 0.80.

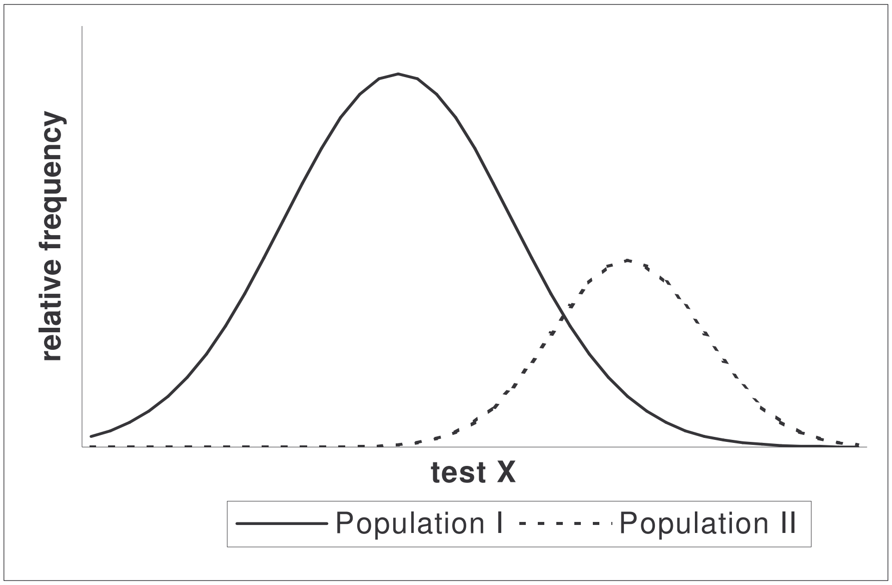


**FIGURE 7.2.** The distribution of test scores within two populations

| | | predictor | | |
|---|---|---|---|---|
| | | - | + | |
| criterion | + | 6 | 14 | 20 |
| | - | 72 | 8 | 80 |
| | | 78 | 22 | 100 |

| | | predictor | | |
|---|---|---|---|---|
| | | - | + | |
| criterion | + | 10 | 40 | 50 |
| | - | 40 | 10 | 50 |
| | | 50 | 50 | 100 |

**FIGURE 7.3.** Classification tables for two base rates

What should we do when missing the diagnosis ‘type-II’ is considered to be worse than the incorrect classification of type-I persons? When the difference between the losses associated with the classification errors is small, the same cut score might be optimal: the cut score might be robust against a small change in the losses. For a larger difference the cut score should be adapted.

Missing a type-II person might be twice as serious as misclassifying a type-I person. In that case we are ready to categorize a subject as a type-II person as long as twice the posterior probability of type-II exceeds the posterior probability of type-I, i.e. as long as $P(\text{II}|X)$ exceeds 1/3. The general classification rule is

$$lP(\text{II}|X) > P(\text{I}|X) \Rightarrow \text{classify as type-II} \tag{7.5a}$$

$$lP(\text{II}|X) < P(\text{I}|X) \Rightarrow \text{classify as type-I,} \tag{7.5b}$$

where $l$ is the ratio of the loss associated with misclassifying a type-II person and the loss associated with misclassifying a type-I person.

When classification errors are serious or when many classification errors are made, one might decide to use test $X$ only as the first screening devise. In that case two cut scores might be used. If a person scores high or low a final classification is made. Scores in between fall in the category *yet undecided* (Cronbach & Gleser, 1965). For other and more difficult classification problems see Hand (1997).

Let us now return to the original example with the classification accepted/rejected and satisfactory/unsatisfactory. Instead of two populations there is only one population. Persons with a criterion score equal to or larger than $Y_c$ are considered to be satisfactory, persons with lower scores are considered unsatisfactory. The utility of the test depends on the cut score between *accept* and *reject*, $X_c$. With a high cut score more persons are incorrectly rejected, with a low cut score more persons are incorrectly accepted. The situation is comparable to that discussed in connection with two populations. In this case decisions must be made on basis of posterior distributions $P(Y|X)$.

Which cut score is optimal depends on the seriousness of the classification errors. To simplify matters, we might consider a discrete loss. Accepting an applicant who is unsatisfactory is equally serious for all those applicants. Rejecting applicants is equally serious for all those applicants incorrectly rejected. One possibility is that incorrectly accepting an applicant is equally serious as incorrectly rejecting an applicant. But it is also possible that one kind of classification errors is considered to be more serious than the other kind of classification errors. We should first determine the losses associated with the two kinds of errors or rather the ratio of these two losses. When we have the loss ratio it is possible to obtain the optimal cut score for a given bivariate distribution of test and criterion scores. For bivariate normally distributed test and criterion scores the optimal cut score is given by Alf & Dorfman (1967). If the two kinds of errors are equally serious, the optimal cut score is the value $X$ for which the expected $Y$ equals $Y_c$, the value of $Y$ on the border between *satisfactory* and *unsatisfactory*. This is easily verified. For this score $X$ the proportion of satisfactory $Y$ equals the proportion of unsatisfactory $Y$ due to the normality of the distribution of $Y$ given score $X$.

It might seem too simple to regard all decisions where persons are incorrectly accepted as equally serious. It might seem more adequate if the seriousness of the classification error depends on the value on the criterion. Van der Linden & Mellenbergh (1977) introduced a linear loss function with decisions on passing and failing examinees. Cronbach & Gleser (1965) gave a systematical treatment of decision making using tests. Petersen & Novick (1976) discussed decisions in the context of culture-fair selection.

In criterion-referenced measurement (Hambleton & Novick, 1973; Popham & Husek, 1969) we are interested in the domain score of an examinee. It is assumed that the examinee has mastered the subject matter if the domain score is at least as high as a standard set on the domain score scale. A test $X$ is used in order to verify whether an examinee has mastered the subject matter. When the test score is high the examinee passes the test, when the test score is low the examinee fails the test. Now, we may look again at Figure 7.1. Instead of criterion $Y$ we have the domain score $\zeta$. Instead of the dichotomies *satisfactory/unsatisfactory* and *rejected/accepted* we have the dichotomies *mastery/nonmastery* and *pass/fail*.

In criterion-referenced measurement it is possible to have a homogeneous population of examinees with a high mastery level. For a reasonable test length reliability is low with such a population. There is a strong regression effect: the expected true score given a low observed score is strongly shifted towards the mean.

As a consequence of the strong regression effect the optimal cut score for ‘pass’ might be relatively low when the average performance on the test exceeds the standard (for standard setting, see Exhibit 7.2). Then the question arises whether we should use the low ‘optimal’ cut score and risk a negative effect on the study commitment of new groups of examinees. In criterion-referenced measurement the problem of low test reliability due to the homogeneity of the population has led to suggestions of alternative coefficients. For example, coefficient kappa (Cohen, 1960) has been suggested as an index for decision consistency (Subkoviak, 1984). It should be noted, however, that decision consistency also might be low when the average mastery level of a homogeneous population is close to the standard of performance.

The computation of coefficient kappa can be illustrated with Figure 7.4. In this figure the crosstabulation is given of the outcomes of two tests. Decision consistency, the proportion of identical decisions, is the sum of the proportions $p_{11}$ and $p_{00}$. In coefficient kappa this proportion is corrected for chance agreement. The coefficient for the two-by-two table is defined as

$$\kappa = \frac{(p_{11} + p_{00}) - (p_{1.}\,p_{.1} + p_{0.}\,p_{.0})}{1 - (p_{1.}\,p_{.1} + p_{0.}\,p_{.0})}\,. \tag{7.6}$$

The decision consistency can be computed with the parallel-test method. In most applications, however, taking a parallel measurement is not practical or possible, and the proportions $p_{11}$ and $p_{00}$ must be estimated from the single administration of a test. The proportions $p_{.1}$ and $p_{.0}$ are set equal to the marginal proportions of the test. Theoretically, the best procedure for the estimation of the proportions $p_{11}$ and $p_{00}$ is the following (see Huynh, 1978):

1. Estimate the distribution of true scores $f(\zeta)$
2. Estimate the conditional error distribution
3. Compute the probabilities $p_{1|\zeta}$ and $p_{0|\zeta}$
3. Compute the proportions $p_{11|\zeta} = p_{1|\zeta} \times p_{1|\zeta}$ and $p_{00|\zeta} = p_{0|\zeta} \times p_{0|\zeta}$
4. Compute $p_{11} = \sum p_{11|\zeta} f(\zeta)$ and $p_{00} = \sum p_{00|\zeta} f(\zeta)$, assuming a discrete true-score distribution.

| | | Test 2 | | |
|---|---|---|---|---|
| | | - | + | |
| Test 1 | + | $p_{10}$ | $p_{11}$ | $p_{1.}$ |
| | - | $p_{00}$ | $p_{01}$ | $p_{0.}$ |
| | | $p_{.0}$ | $p_{.1}$ | 1.00 |

**FIGURE 7.4.** Decisions on two parallel tests

**Exhibit 7.2. Standard Setting for Performance**

In educational and psychological assessment the uses and interpretations of standards are essential, hence attention has to be paid to standard setting methods for test performance. Following Hambleton (1996, p. 919) performance standards may be defined as the scores, or profiles of scores that must be achieved by examinees to be classified as, say, proficient. Consequently, a critical step in the development and use of these assessment procedures is to establish cut points dividing the score range into categories that are meaningful for the educational or psychological community (cf. *Standards*, APA, AERA, & NCME, 1999, p. 53). An example is how to classify students on the basis of their score on NAEP Maths Assessments into categories (a) basic, (b) proficient, and (c) advanced.

It depends on the modeling of the classification problem how the standard-setting method turns out in practice. A plethora of standard-setting methods has been developed (see e.g. Berk, 1986; De Gruijter, 1985 on compromise models; Hattie, Jaeger, & Bond, 1999 and the references mentioned there; Hambleton, 1996; Cizek, 2001) leading to different results (Jaeger, 1989). Nor is it likely that repeated application of the same method in connection with different tests gives equivalent results. Standard setting is an important and inevitable activity, but also an activity that remains based on the best subjective judgement of experts.

The oldest procedure to set a standard is the Nedelsky method (Nedelsky, 1954) for multiple choice items. The procedure is as follows: one imagines a hypothetical borderline examinee, an examinee whose performance level marks the transition between satisfactory and unsatisfactory performance. For each item one decides which distractors this examinee can eliminate as incorrect; the probability of a correct response to an item is given as the inverse of the number of remaining options. The sum of the probabilities gives the standard of performance.

When a new test version is introduced, a standard for this test version must be set. There are several possibilities to set the standard on a new test form. When the old test and the new test have items in common, one of the procedures for test equating (Chapter 10) might be applied.

## 7.5 Selection and Classification with More Than One Predictor

When the quality of the classification or diagnosis based on a single measurement $X$ is too low, other measurement instruments should be considered for inclusion in a test battery used for selection and classification purposes. What characteristics should potential additional tests have? In general it seems wise to select tests that give new information useful for making the classification or diagnosis. So, added tests should

a. correlate with the criterion
b. have a low correlation with other predictors.

We will first demonstrate the point using a classical procedure of an unweighted sum of predictor scores. We will use the argument that has been used by Gulliksen (1950) with respect to the selection of items for a test. The validity of the sum of $n$ scores $X_i$, $X$, with respect to criterion $Y$ can be written as

$$r_{XY} = \frac{\sum_{i=1}^{n} s_Y s_i r_{iY}}{s_Y \sum_{i=1}^{n} s_i r_{iX}} = \frac{\text{ave}(s_i r_{iY})}{\text{ave}(s_i r_{iX})}, \tag{7.7}$$

where $s_i$ is the standard deviation of the scores on test $X_i$. Let us assume that the variances of the tests do not differ much. In that case a test that highly correlates with the criterion and not so much with the other predictors, adds to the numerator and not to the denominator. When we select items for a test in order to maximize validity, items are selected that may decrease the reliability of the test.

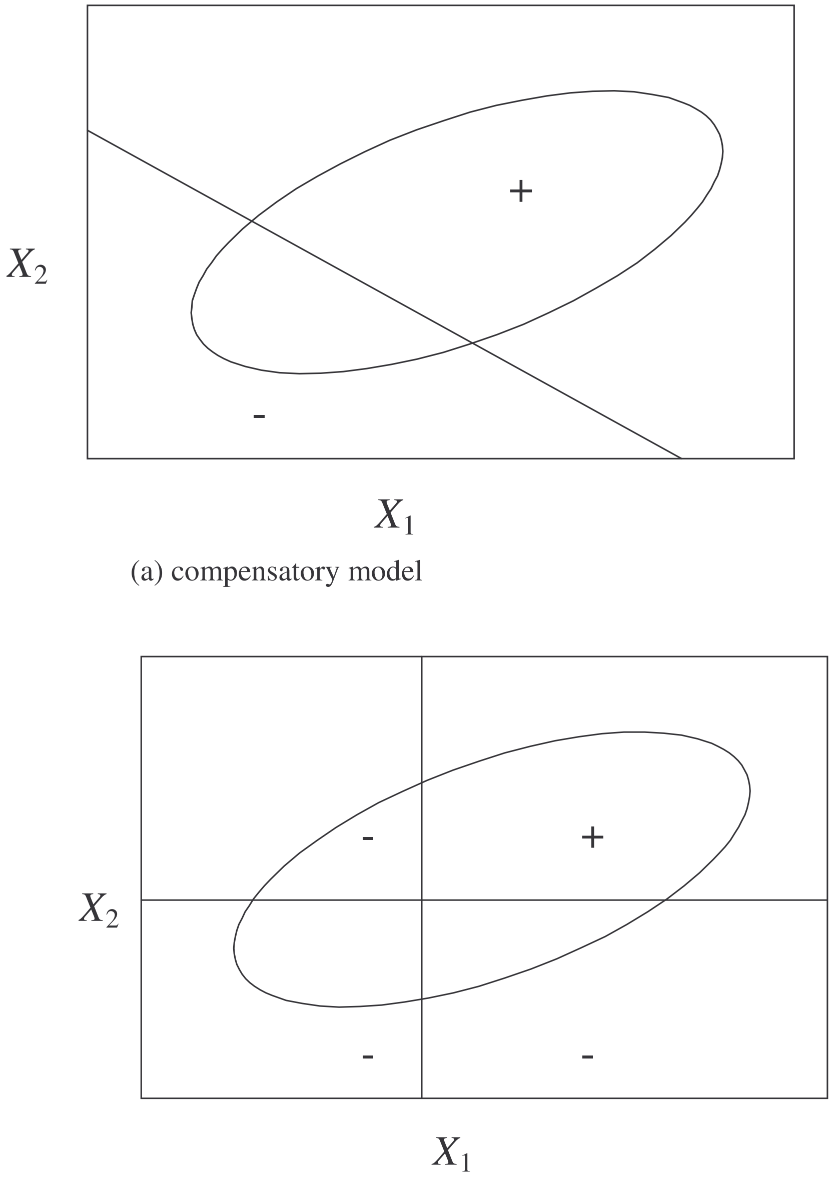
The unweighted sum of predictor scores does not give the optimal combination of measurements. The obvious method is to use a weighted combination of scores. Optimal weights can be obtained from a multiple-regression analysis, where optimality is operationalized in terms of a least-square loss function. With scores on two predictors $X_1$ and $X_2$, and a criterion $Y$ the formula for the regression has the form

$$\hat{y}_p = a_0 + a_1 x_{1p} + a_2 x_{2p} \,. \tag{7.8}$$

The linear regression approach exemplifies the so-called compensatory model for selection. In this model the minimum requirement on the criterion is achieved by an additive combination of abilities. A low level of achievement for one ability can be compensated by a high level for another ability. The compensatory model for selection with two predictors is displayed in Figure 7.5.a. With errorless variables $X_1$ and $X_2$ the straight line gives all combinations of $x_1$ and $x_2$ that result in the critical criterion level. So, the straight line is the border between combinations of abilities $x_1$ and $x_2$ that correspond to a satisfactory criterion level (+), and combinations $x_1$ and $x_2$ that correspond to an unsatisfactory criterion level (-). The linear regression formula is adequate for the compensatory model of Figure 7.5.a. The linear regression formula can be adapted to a degree of nonlinearity in the relation between the criterion and the predictors: powers of the predictor scores ($x_1^2$, etc.) can be added as predictors in the multiple regression formula.

Two other models may be discussed in the context of selection: the conjunctive model and the disjunctive model (Coombs, 1964). The conjunctive model requires persons to satisfy a minimum level of achievement on each of the relevant abilities. There is no possibility of compensation. The conjunctive model with two abilities $X_1$ and $X_2$ is illustrated in Figure 7.5.b. The conjunctive model seems to ask for multiple cut scores as in the figure. Actually, the classification of examinees in the conjunctive model is more complicated than that when the predictors have reliabilities less than one. Then the score on one measurement instrument contains information with respect to the achievement on another measurement instrument, assuming that the measurement instruments are correlated. Lord (1962) demonstrated that the prediction of criterion performance is increased when some amount of compensation between the fallible measurements is allowed.

In the third model, the disjunctive model, persons may satisfy the criterion by having at least one sufficient ability. With two abilities $X_1$ and $X_2$ only the quadrant with both low $X_1$ and low $X_2$ is associated with unsatisfactory criterion performance.



(a) compensatory model

(b) conjunctive model

**FIGURE 7.5.** Classification with two predictors

## 7.6 Convergent and Discriminant Validation: A Strategy for Evidence-Based Validity

Evidence for construct-related validity can be obtained in quite a number of ways, i.e., using several research designs and corresponding statistical methods for data analysis. These methods in construct validation, proposed already by Cronbach & Meehl (1955) and more specifically by Campbell & Fiske (1959), may be summarized as follows:

- the study of group differences

If we expect two or more groups to differ on the test purportedly measuring a construct, this expectation may be tested directly resulting in evidence for construct-related validity.

- the study of correlations between tests and factor analysis
  If two or more tests are presumed to measure some construct, than a factor analysis of the correlation matrix must reveal one underlying factor as an indicator of the common construct.

- studies of internal structure
  For many constructs, evidence of homogeneity within the test is relevant in judging validity.

- studies of change over occasions
  The stability of test scores, i.e. retest reliability, may be relevant to construct validation.

- studies of process
  Observing a person's process of performance is one of the best ways of determining what accounts for variability on a test (see e.g. Cronbach & Meehl, 1955; Cronbach, 1990). In addition, judgment and logical analysis are recommended in interpretations employing constructs (cf. Cronbach, 1971, p. 475).

An important elaboration and extension of the study of correlations between tests is Campbell & Fiske's (1959) so-called convergent and discriminant validation by the multitrait-multimethod matrix.

*The Multitrait-Multimethod Approach*

Suppose we have a few different measurement instruments of a trait. We might expect that these measurements correlate. If there really is an underlying trait the correlations should not be too low. On the other hand, correlations between measurement instruments for different traits should not be too high; otherwise it makes no sense to make a distinction between the traits. In many investigations several traits are measured using the same kind of instrument, for example a questionnaire. This might pose a problem. When there is a correlation between two traits measured with the same method, we might wonder to which extent the correlation is due to the covariation of the traits and to which extent it is due to the use of a single measurement method. Campbell & Fiske (1959) proposed to use the multitrait-multimethod matrix research design in order to study the convergence of trait indicators and the discriminability of traits in validation studies. A hypothetical example with three traits and three methods is given in Figure 7.6.

| | | | method a | | | Method b | | | method c | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | trait 1 | trait 2 | trait 3 | trait 1 | Trait 2 | trait 3 | trait 1 | trait 2 | Trait 3 |
| method a | trait 1 | $r_{11}(aa)$ | $r_{12}(aa)$ | $r_{13}(aa)$ | $r_{11}(ab)$ | $r_{12}(ab)$ | $r_{13}(ab)$ | $r_{11}(ac)$ | $r_{12}(ac)$ | $r_{13}(ac)$ |
| | trait 2 | | $r_{22}(aa)$ | $r_{23}(aa)$ | $r_{21}(ab)$ | $r_{22}(ab)$ | $r_{23}(ab)$ | $r_{21}(ac)$ | $r_{22}(ac)$ | $r_{23}(ac)$ |
| | trait 3 | | | $r_{33}(aa)$ | $r_{31}(ab)$ | $r_{32}(ab)$ | $r_{33}(ab)$ | $r_{31}(ac)$ | $r_{32}(ac)$ | $r_{33}(ac)$ |
| method b | trait 1 | | | | $r_{11}(bb)$ | $r_{12}(bb)$ | $r_{13}(bb)$ | $r_{11}(bc)$ | $r_{12}(bc)$ | $r_{13}(bc)$ |
| | trait 2 | | | | | $r_{22}(bb)$ | $r_{23}(bb)$ | $r_{21}(bc)$ | $r_{22}(bc)$ | $r_{23}(bc)$ |
| | trait 3 | | | | | | $r_{33}(bb)$ | $r_{31}(bc)$ | $r_{32}(bc)$ | $r_{33}(bc)$ |
| method c | trait 1 | | | | | | | $r_{11}(cc)$ | $r_{12}(cc)$ | $r_{13}(cc)$ |
| | trait 2 | | | | | | | | $r_{22}(cc)$ | $r_{23}(cc)$ |
| | trait 3 | | | | | | | | | $r_{33}(cc)$ |

**FIGURE 7.6.** The multitrait-multimethod correlation matrix with three methods and three traits

The main diagonal contains the reliabilities; we might call these entries monotrait-monomethod correlations. In the first diagonal entry, for example, we have $r_{11}(aa)$, the reliability of the measurement instrument which measures trait 1 by means of method $a$. Adjacent to the main diagonal we have triangles with heterotrait-monomethod correlations. We also have blocks with correlations involving two different methods. Within these blocks we have diagonals with correlations involving one trait. These monotrait-heteromethod values are the so-called validity diagonals; a gray background in the figure indicates the monotrait-heteromethod entrees.

According to Campbell and Fiske a validation process is satisfactory if:

1. Correlations between measurements of the same trait with different methods are significantly larger than 0. Then we have convergence.
2. Correlations between measurements of a trait with different methods are higher than the correlations of different traits measured with the same method. The validity diagonals should be higher than the correlations in the monomethod-heterotrait triangles. In that case we have discriminant validity.
3. A validity coefficient $r_{ii}(ab)$ is larger than the correlations $r_{ij}(ab)$ and $r_{ji}(ab)$.
4. In the heterotrait triangles of the monomethod blocks and the heteromethod blocks the pattern of correlations is the same.

Campbell & Fiske (1959) considered only informal analysis and eye-balling techniques for the study of multitrait-multimethod matrices. Such matrices, however, may also be analyzed with generalizability theory (Cronbach et al., 1972). An alternative approach is to use confirmatory factor analysis. It belongs to the class of structural equation modeling (SEM), and is, among others, a promising procedure to obtain evidence of construct-related validation where more constructs are involved in a nomological network. Also, with more than one measure, confirmatory factor analysis with so-called structured means can be used to test hypotheses with respect to the tenability of equivalence conditions (e.g. strictly parallel measures, tau-equivalent measures) of a set of measures. Last but not least, this type of confirmatory factor analysis offers a fruitful approach to test validation. Technical details of SEM can be found in Jöreskog (1974), more general details on alternative approaches for an analysis of multitrait-multimethod matrices can be found in Schmitt, Coyle & Saari (1977) and Schmitt & Stults (1986) while Byrne (1994) gives procedures and examples for testing the factorial validity of a set of measures as well as construct validity. The reader should, however, be warned: routine applications of SEM for the analysis of multitrait-multimethod matrices are doomed to fail due to all the pitfalls in the use of SEM. The lesson from all this is that none of the analytic approaches to multitrait-multimehod matrices should be done routinely. A thoughtful and well-balanced review of approaches to the multitrait-multimethod matrix is recently given by Crano (2000).

Test manuals should provide information on reliability, validity and test norms. The manuals cannot be exhaustive, however. After publication of a test new research adds to the validation of test uses. Summaries of research and critical discussions of tests are needed. The Mental Measurement Yearbooks fulfil such a function. Let us take the Beck Depression Inventory, a frequently cited inventory. The BDI is reviewed by two reviewers in the Thirteenth Mental Measurement Yearbook, Carlson (1998) and Waller (1998). The BDI is a brief self-report inventory of depression symptoms. It has 21 items, scored on a 4-point scale. The test is used for psychiatric patients. It also is frequently used as a screening device in normal populations. The manual gives information on reliability, validity and test norms. But, the reviewers argue that the manual is too short. Much useful information must be found in other published sources. Several aspects of validity are discussed by the reviewers. The inventory has face validity: the items are transparent. The high face validity makes the inventory vulnerable to faking. Correlations with other tests have been computed and a factor analysis has been done. The inventory discriminates patients from normal persons. Waller notes that the information with respect to discrimination validity is lacking. What is, for example, the correlation of the BDI with an anxiety measure, a measure of a different construct?

### 7.7 Validation and IRT

Item response theory (IRT) provides models in which the responses of subjects on the individual test items are modeled. IRT models not only allow for the estimation of person and item parameters, but also a statistical test for how good the model fits the data. So when a unidimensional IRT model is assumed, the test of model fit informs us about the existence of a single construct or latent trait underlying the observed item responses. IRT models are discussed in later chapters, so here it suffices to state in general terms the nature of the construct-related validation using IRT. To date, this is a promising terrain of research, more can be found in Embretson & Prenovost (1999) and the references mentioned there.

Considering construct-related validity in the context of IRT does not exhaust the validity issue in psychometrics at large. In the next section research validity will be discussed in the broad context of empirical behavioral research.

### 7.8 Research Validity: Validity in Empirical Behavioral Research

Empirical behavioral research is a broader context of research than test R&D. Therefore, more general aspects of validation are involved, and the type of validity in this broader context has been coined research validity (see e.g. Judd & Kenny, 1981).

The tenability of theories and the generalizability of findings from empirical research are influenced by four classes of validation:

1. statistical conclusion validity
   This is defined as the extent to which the design of the study is sufficiently sensitive or powerful to detect outcome effects. It addresses the question whether the relationship(s) observed in the sample are due to chance or not.
2. internal validity
   This is the extent to which detected outcome effects, viz. test scores, are due to the operationalized cause rather than to other rivaling causes. The question of

internal validity is often rephrased as: are there no alternative explanations for the detected effects in terms of e.g. changes in test scores.

3. external validity
This is defined as the extent to which the detected outcome effects (test scores for that matter) can be generalized to theoretical constructs, subjects, occasions, and situations other than those specified in the original study. In most instances, this type of validity is used to refer to the question whether the test scores or other effect measures that were found in the sample can be assumed to exist in the population (or, a certain, well-defined population) as well.

4. construct validity
This type of validity is defined similarly as in the trichotomy above, as the extent to which the theoretical constructs in a study have been successfully operationalized. In other words, does the measurement on a certain variable represent the phenomenon or propensity it is supposed to measure.

These and similar definitions of research validities can be found in textbooks on research methodology. Elaborations have been proposed; a succinct overview is given by Cook & Shadish (1994; see also Cook, Campbell & Peracchio, 1990).

Although research validities are generally not to be considered as part of test theory per se, it is important and relevant to point out that each of the above four types of research validity may be under threat of one sort or another. Among these threats are history, maturation, testing, instrumentation, statistical regression towards the mean, mortality. One approach to circumvent one or more of these threats is choosing the appropriate design of the study and proceed along the road of performing a generalizability study. So here we see that the demarcation of psychometric reliability and validity is blurred. Earlier, generalizability studies have been treated as a liberalization of classical test reliability. It also serves as a vehicle for validation studies.

## Exercises

7.1 In a study test $X$ is administrated to all persons. Test $Y$ is administrated to a selection of persons. Within each group with the same score on $X$ persons are randomly chosen for selection into the group that is administered test $Y$. The correlation between $X$ and $Y$ equals 0.8. The variance of the scores on $X$ within the selection equals 36.0. The variance on $X$ in the total group equals 16.0. Estimate the correlation between $X$ and $Y$ in the total group.

7.2 Given is a 10-item test with the following frequency distribution in two groups A and B:

| Score | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $F_A$ | 0.043 | 0.109 | 0.130 | 0.174 | 0.217 | 0.174 | 0.087 | 0.043 | 0.022 | 0.0 | 0.0 |
| $F_B$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.045 | 0.091 | 0.136 | 0.182 | 0.227 | 0.182 | 0.136 |

We want to use the test in order to classify persons in the future. Both kinds of errors are equally serious. At what test score should we take the decision to classify a person as a 'B-person' assuming that population A has four times the size of population B? Can you comment on your result?

7.3 What happens if the base rate of belonging to group B in exercise 7.2 is 0.5 instead of 0.2?

7.4 We have a test with a mean equal to 75.0, a standard deviation equal to 8.0 and a reliability equal to 0.25. With the test we want to decide which examinees are masters and who are nonmasters. The criterion of mastery is 70.0 on the true-score scale. The errors of classifying masters and nonmasters incorrectly are equally serious. Compute the optimal cut score under the assumption that the observed scores and true scores have a bivariate normal distribution.

7.5 Compute coefficient κ for the data in the following table:

| | | |
|---|---|---|
| + | 10 | 60 |
| - | 20 | 10 |
| | - | + |

# 8 ITEM RESPONSE MODELS

## 8.1 Introduction

Item response theory is a general term for a family of models, the item response or IRT models, that share some fundamental ideas. These ideas are that IRT models persons' responses on individual items. The response of a person on a test item is conceived of as a function of a person characteristic and an item characteristic. The response of a person, i.e. the performance of an examinee, is assumed to depend upon one or more factors called (latent) traits or abilities. Each item of a set of items measures the underlying trait or traits. An example of a simple IRT model is that a person's performance on an item depends only on one underlying trait, and that the relationship between persons' performance on an item and the trait underlying item performance can be described by a monotonically increasing function. The latter function is commonly called *item trace line*, *item characteristic function* (ICF) or *item characteristic curve* (ICC). It specifies how the probability of a correct response to an item increases as the level of the trait increases. In contrast to classical test theory and generalizability theory discussed earlier, IRT consists of a class of mathematical models for which estimation procedures exist for model parameters (i.e. person and item parameters) and other statistical procedures for investigating to what extent the model at hand fits the data or persons' responses to a set of items.

IRT research and developments not only pervade scholarly journals, in the latest edition of the Standards of psychological and educational testing (APA, AERA, & NCME, 1999) also ample space is given to IRT. Ability or trait parameters, counterparts of classical true scores and G theory universe score, are mentioned, whereas summarizing IRT-based test information functions is also recommended (See Section 8.13).

In Section 8.2 the basic concepts of IRT will be discussed, and several unidimensional models for dichotomous data will be introduced. Apart from the types of IRT models in terms of a specification of the ICC, models can also be distinguished as to the number of response options modeled, and also more than one latent trait can be postulated, leading to multidimensional IRT models (see also 8.2 and 8.3). In Section 8.4 item-test regression will be considered, and compared to IRT item-trait regression. It has already been said that IRT leads to the estimation of model parameters; the estimation of item parameters is introduced in Section 8.5. In Section 8.6 the joint estimation procedure (JML) for item as well as person parameters is discussed. To what JML leads in the Rasch model can be found in Section 8.7. Other estimation methods with their characteristic properties will be discussed in Section 8.8 (the marginal maximum likelihood or MML method) and in Section 8.9 (the conditional maximum likelihood or CML method for the Rasch model). In Section 8.10 some specific problems will be discussed with respect to the estimation of item parameters. Section 8.11 is on maximum likelihood estimation of person parameters. This does not exhaust the possibilities for the estimation of person parameters; in Section 8.12 Bayesian estimation is mentioned.

The IRT concepts of item information and test information break away from the concept of the variance of measurement errors being constant over the whole range of scores. These information concepts are elaborated in Section 8.13. As IRT gives a model-approach to measurement, model fit is also a central theme (see 8.14). Finally, for the interested reader maximum likelihood estimation in the context of the Rasch model is discussed in an appendix (Section 8.15).

## 8.2 Basic Concepts

In a unidimensional model we assume that the responses can be described by a model with one latent dimension, i.e. as if only one dimension ability or latent trait accounts for the responses. The latent ability θ is defined in most models on a scale of minus infinity to plus infinity (-∞, ∞). The probability of a particular response to a dichotomous item is a monotonous and nonlinear function of the ability. The probability of a correct response increases with increasing ability or latent trait value (for an exception in connection with a model involving guessing behavior see Samejima, 1979). The conditional probability correct, the probability of a correct response given ability, might be interpreted as the probability of a correct response for a randomly selected person with the given ability (Holland, 1990).

The probabilities are bounded by the values 0 and 1. The assumption of *unidimensionality* implies that the responses to different items are independent given the latent trait. We have obtained *local independence*. If we did not have local independence, one dimension would not be enough to account for the responses. In that case local independence would be obtained given all relevant latent abilities.

*The Rasch Model*

In Figure 8.1 the probability of a correct response as a function of latent ability θ is given for two hypothetical items. The two items are Rasch items; i.e. they satisfy the Rasch model assumption of equal discriminability for items with correct-incorrect scoring. That is to say, Rasch curves are parallel.

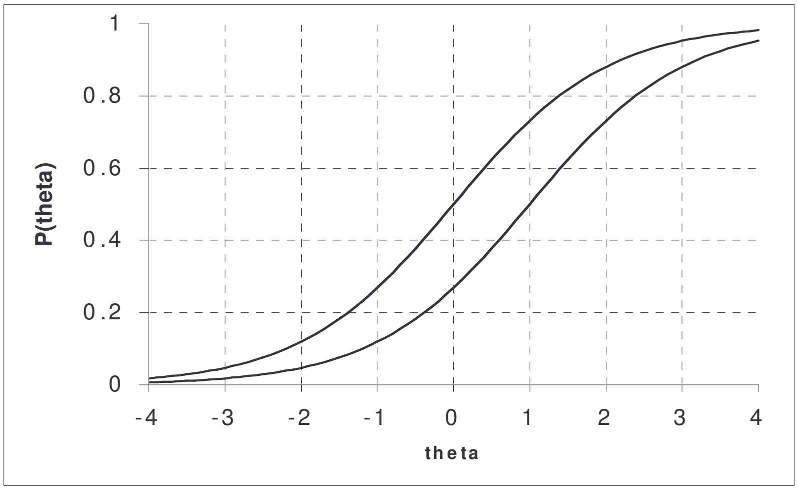


**FIGURE 8.1.** Item characteristic curves for two Rasch items

In the Rasch model the probability of a correct response on item *i*, given ability or person parameter θ, is equal to

$$P_i(\theta) = P(x_i = 1 \mid \theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} = \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)}, \tag{8.1}$$

where $b_i$ is the item difficulty parameter of item *i* (Rasch, 1960). The probability varies from 0.0 for θ = -∞ to 1.0 for θ = ∞. Most of the variation in probability lies in the interval from θ

$=b_i - 4.0$ to $\theta = b_i + 4.0$. For $\theta$ equal to $b_i$ the probability equals 1/2. In Figure 8.1 the curve on the left is the curve of a Rasch item with $b_i = 0$, the curve on the right belongs to a Rasch item with $b_i = 1.0$.

The person parameter $\theta$ and the item parameter $b$ occur in (8.1) only in the combination $\theta$ - $b$; When we take log-odds, the natural logarithm of the ratio of the probability of a correct response and the probability of an incorrect response, we obtain the logit:

$$\ln\left(\frac{P_i(\theta)}{1-P_i(\theta)}\right) = \theta - b_i\,.$$

We conclude from this that the probability of a correct response (8.1) remains invariant if we increase the value of $\theta$ to $\theta^* = \theta + d$ and simultaneously increase the value of $b$ to $b^* = b + d$. The parameters of the Rasch model are defined on an additive scale, a special case of the interval scale. The consequence is that in an application of the model always one restriction on the parameters is needed in order to fix the latent scale. We might, for example, set the item parameter of one of the items equal to 0.0. Another possible restriction is to set the mean of the item parameters equal to 0.0.

*Two- and Three-Parameter Logistic Models*

The Rasch model is a one-parameter model. The model has one item parameter: the item difficulty parameter. In many tests the items not only differ with respect to difficulty but also with respect to discriminating power. The two-parameter *logistic model* (Birnbaum, 1968), 2PL model for short, has a second item parameter, the item discrimination parameter. The model is given by the equation

$$P_i(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{8.2}$$

where $a_i$ is the discrimination parameter of item $i$. The term logistic refers to the fact that the right hand side of equation (8.2) is equal to the cumulative logistic distribution function. The slope of the ICC at $\theta$ is equal to $a_i P_i(\theta)[1 - P_i(\theta)]$. At $\theta = b_i$ the slope is equal to $0.25a_i$. So the slope at $\theta = b_i$ is steeper for higher values of $a_i$.

The person parameters in the 2PL model are defined on an interval scale. The probability of a correct response does not change if we transform $\theta$ into $\theta^* = d\theta + e$ under simultaneous transformations $b^* = db + e$ and $a^* = a/d$. In order to fix the latent scale we need two restrictions. We might, for example, set the mean $\theta$ equal to 0.0 and the standard deviation of the $\theta$'s equal to 1.0.

In Figure 8.2 two item characteristic curves are displayed, one with $a_i = 1.0$, the other with $a_i = 10.0$. One can imagine what will happen to the ICC if the discrimination parameter of an item increases indefinitely. The item characteristic curve approximates a jump function with a value equal to 0 for $\theta$ smaller than $b_i$, and a value equal to 1 for $\theta$ larger than $b_i$. We then have a Guttman item with a perfect discrimination at the value $\theta = b_i$, and no discriminating power to the left and to the right of this point.

The item characteristic curves in Figure 8.2 cross. In the Rasch model item characteristic curves do not cross, but run parallel. The Rasch model is not the only probabilistic model with nonintersecting item characteristic curves. Another model with this property is the nonparametric Mokken model of double monotonicity (Mokken, 1971).

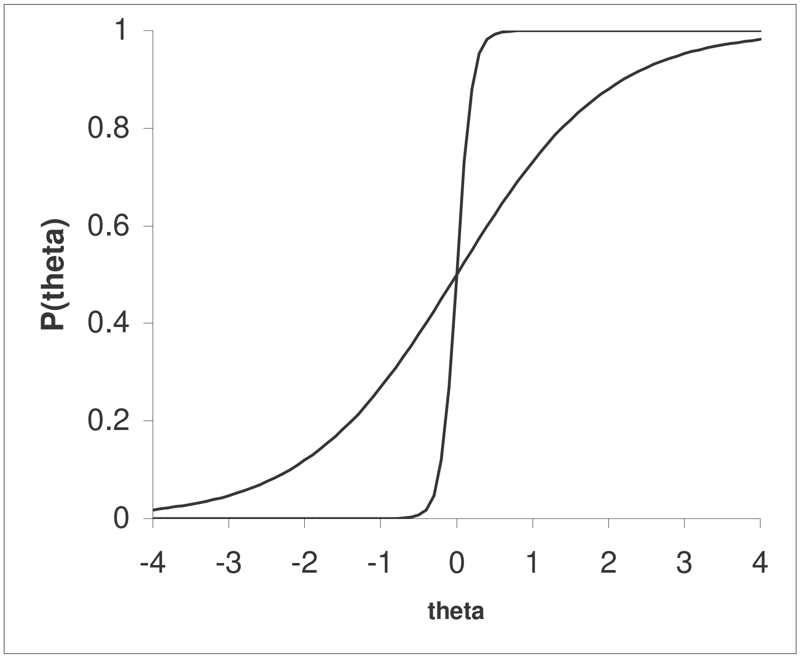


**FIGURE 8.2.** Item characteristic curves for two items with different discrimination parameters

With items of the multiple-choice type guessing is possible and cannot be excluded. If an examinee does not know the answer on a four-choice item, he or she might correctly guess the answer with a probability equal to 1/4. With this kind of items one better introduces a lower asymptote larger than 0 for the item characteristic curve. So one obtains the three-parameter logistic model

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{8.3}$$

where $c_i$ is the lower asymptote. The third parameter also is called the *pseudo-chance-level parameter*. This parameter is not set equal to the inverse of the number of response alternatives but it is estimated along with the other item parameters. Figure 8.3 displays two items, one with $c_i$ equal to 1/4, the other with $c_i$ equal to 0.0. The influence of the third item parameter at the lower level of θ is clear.

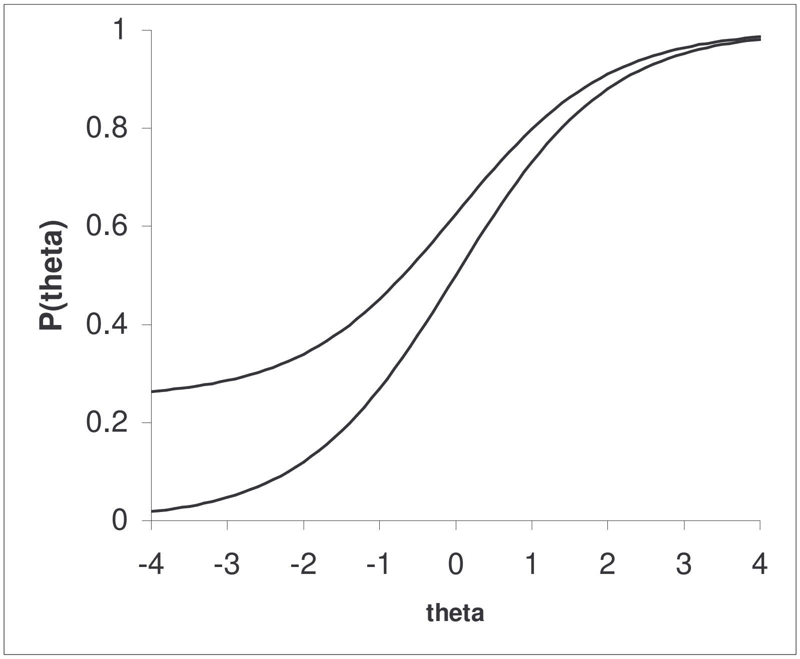


**FIGURE 8.3.** Item characteristic curves for two items with different pseudo-chance-level parameters

The 2PL model is a special case of the 3PL model with $c_i = 0.0$ for all items. A 1PL model is obtained if all item discrimination parameters are set equal. In the Rasch model (8.1) this common discrimination parameter is set equal to 1.0. The differences between the models seem to be very clear. Meredith and Kearns (Meredith & Kearns, 1973), however, demonstrated that a special case of the 3PL model can be reformulated in terms of the Rasch model.

Besides the logistic model we have the *normal ogive model*. It has an ICC with the form of the cumulative normal distribution. This model was the first to be used in test theory (see Lord, 1952). The two-parameter normal ogive model is given by

$$P_i(\theta) = \Phi[a_i(\theta - b_i)] = \int_{-\infty}^{a_i(\theta-b_i)} \varphi(t)dt = \int_{-\infty}^{a_i(\theta-b_i)} \frac{1}{\sqrt{2\pi}} \exp[-\tfrac{1}{2}t^2]dt \,. \tag{8.4}$$

The normal ogive plays a role in some models with more dimensions and/or more than two response categories (Muthén, 1984; Bock, Gibbons & Muraki, 1988; Muraki & Carlson, 1995). The application of the model to polytomous, multidimensional data will be discussed in Section 8.3.

The normal ogive model and the logistic model give practically the same probabilities if a scaling factor $D$=1.7 is introduced in the logistic model:

$$\Phi[a_i(\theta - b_i)] \approx \frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]} \,.$$

For this reason the factor $D$ frequently is part of the logistic model as described in the literature. If the parameters of the logistic model are given by (8.2) or (8.3), the parameters are defined in the ‘logistic metric’. They can be transformed to the ‘normal metric’ through a division of the discrimination parameters by the scaling factor 1.7.

*Other IRT Models*

Fischer (1983) extended the Rasch model with linear constraints on the item parameters. Other extensions of the models have been proposed: extensions to more than two response categories and extensions to more than one latent trait dimension. Bock (1972) proposed a general *nominal response model*. In this model the probability of a response to item option *k* from *m* available options is given by

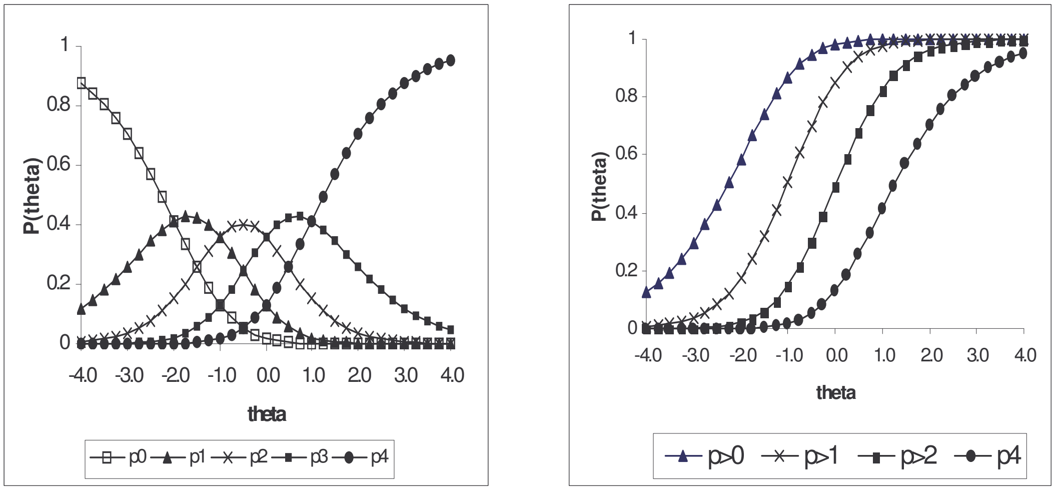
$$P_{ik}(\theta) = \frac{\exp[a_{ik}(\theta - b_{ik})]}{\sum_{h=1}^{m} \exp[a_{ih}(\theta - b_{ih})]}. \tag{8.5}$$

Extensions to items with more than two categories have been formulated notably for the Rasch model (e.g. Andersen, 1977; Andrich, 1978; Masters, 1982). Two closely related Rasch models for items with more than two categories are well known: the *rating scale model* (Andrich, 1978, 1999) and the *partial credit model* (Masters, 1982, 1999). The rating scale model can be obtained as a submodel of the partial credit model by a reparametrization of the parameters. The models differ in their substantive interpretations, however.

The partial credit model models partial understanding in problems with multiple steps. In the partial credit model with categories 0, …, *m* the probabilities for categories other than category 0 are given by

$$P_{ik}(\theta) = \frac{\exp[\sum_{j=1}^{k} (\theta - \delta_{ij})]}{1 + \sum_{h=1}^{m} \exp[\sum_{j=1}^{h} (\theta - \delta_{ij})]}. \tag{8.6}$$

An exemplary item is given in Figure 8.4.



(a) option characteristics curves (b) probability of exceeding a specific level

**FIGURE 8.4.** The partial credit model

The model parameters $\delta_{ij}$ are "step" difficulties governing the "step" probabilities $P_{ik}/(P_{ik}+P_{i,k-1})$, which have the form of Rasch model items. The model is, for example, applied when several dichotomous items are related and form an item cluster. Huynh (1994) demonstrated the applicability of the PCM model for a testlet composed of independent Rasch items. When the steps of an item are scored sequentially (all steps after the first error in a chain of steps are evaluated as failed), the model is not suitable. Sequentially scored items must be modeled by a set of binary items with missing observations after the first failure (Akkermans, 2000).

In the rating scale model thresholds are modeled for the different categories in items with ordinal response categories (for example, *poor*, *fair*, *good*, *excellent*). In the rating scale model for $m + 1$ score categories (0, . . ., $m$) the probability of choosing category $k$ of item $i$ can be written as

$$P_{ik}(\theta) = \frac{\exp[k(\theta - b_i) - \sum_{j=0}^{k} \tau_{j(m+1)}]}{\sum_{h=0}^{m} \exp[h(\theta - b_i) - \sum_{j=0}^{h} \tau_{j(m+1)}]}, \tag{8.7}$$

where $\tau_{j(m+1)}$ are threshold parameters for all items with a common number of $m + 1$ categories, with $\tau_{0(m+1)} = 0$.

Samejima (1969) published on a *graded response model* for items with ordered response options. For the logistic model the probability of choosing option $k$ or a higher option equals

$$P^*_{ik}(\theta) = \frac{\exp[a_i(\theta - b_{ik})]}{1 + \exp[a_i(\theta - b_{ik})]}, \tag{8.8}$$

where the option parameter $b_{ik}$ increases with $k$. In this model the probability of option $k$ as a function of latent ability, the option characteristic function, is

$$P_{ik}(\theta) = P^*_{ik}(\theta) - P^*_{i,k+1}(\theta), \tag{8.9}$$

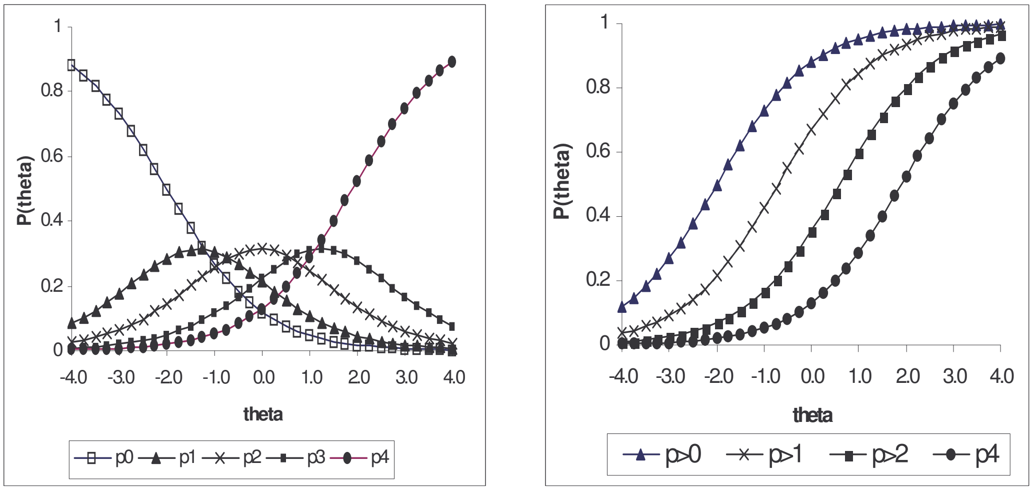
with $P^*_{i0}(\theta) = 1.0$, $P^*_{i,m+1}(\theta) = 0.0$. An example of an item in the graded response model is given in Figure 8.5. The probability of category 0 decreases with $\theta$, the probability of category $m$ increases with $\theta$, and the probabilities of the intermediate categories have their peaks in the order of the categories. A submodel of this model is the model where $b_{ij}$ can be partitioned into an item parameter $b_i$ and a scale parameter $\tau_j$ (Muraki, 1990).

In the graded response model the probability of a response in category $k$ or $k + 1$ is

$$\begin{aligned} P_{ik}(\theta) + P_{i,k+1}(\theta) &= [P^*_{ik}(\theta) - P^*_{i,k+1}(\theta)] + [P^*_{i,k+1}(\theta) - P^*_{i,k+2}(\theta)] \\ &= P^*_{ik}(\theta) - P^*_{i,k+2}(\theta), \end{aligned}$$

which has the same form as the original probabilities. The response categories *poor* and *fair*, and the response categories *good* and *excellent* of the four response categories *poor*, *fair*, *good*, *excellent* might, for example, be combined for the analysis of the response data. So, the graded response model allows a dichotomization of the response categories, and – at the cost of loosing some information - an analysis with an IRT-model for dichotomous data.

Combining categories is not possible with the Rasch models for polytomous data without violation of the model (Jansen & Roskam, 1986). So, in case combining of categories is an obvious possibility with the data that are to be analyzed with an IRT-model, the graded response model is the more realistic model despite the statistical advantages of the polytomous Rasch models. However, the partial credit model might give comparable results.



(a) option characteristics curves (b) probability of exceeding a specific level

**FIGURE 8.5.** The graded response model

Molenaar has presented a generalization of the Mokken model to the case of more than two categories (Molenaar, 1997). Ramsay (1991) suggested a general approach to nonparametric estimation. An introduction to nonparametric modeling is given by Molenaar and Sijtsma (Molenaar & Sijtsma, 2002). Rasch developed a Poisson model (see also Lord & Novick, 1968) for the number of mistakes in a test. Models for speeded and time-limit tests can be found in Roskam (1997).

The other extension is to more dimensions. Item response models actually are factor-analytic models with a nonlinear relationship between factor and expected scores. Most models that have been proposed are compensatory: a low ability *i* can be compensated by a high ability *j* (like in (4.12)). A factor-analytic approach to a multidimensional latent space has been given by Bock, Gibbons & Muraki (1988), McDonald (1997, 1999) with NOHARM (Fraser, 1988), Muraki & Carlson (1995), Reckase (1997) and Muthén (1984) with LISCOMP. Shi & Lee (1997) discussed the estimation of latent abilities for the nonlinear factor model. For multidimensional Rasch models, see Adams, Wilson & Wang (1997). A non-compensatory multidimensional model in which cognitive processes are modeled has been proposed by Embretson (1984).

## 8.3 The Multivariate Normal Distribution and Polytomous Items

In the factor analytic model the score of person *p* on item *i* can be written as

$$Z_{ip} = \alpha_{i1}\theta_{1p} + \alpha_{i2}\theta_{2p} + \ldots + \alpha_{in}\theta_{np} + E_{ip}. \tag{8.10}$$

The additive constant has been dropped from the equation. This can be done without consequences for the generality of our argument.

In this chapter the model is defined in terms of latent variables. The $Z_{ip}$ are not observed. These variables describe response processes. When $Z_{ip}$ exceeds a certain threshold, a score in a given category of polytomous item $i$ is observed. For item $i$ with $m + 1$ categories we have

$$\begin{aligned} X_{ip} &= 0 \text{ if } Z_{ip} < \gamma_{i0} \\ X_{ip} &= k \text{ if } \gamma_{i,k-1} \leq Z_{ip} < \gamma_{ik} \\ X_{ip} &= m \text{ if } \gamma_{i,m-1} < Z_{ip}. \end{aligned} \tag{8.11}$$

Let us write $P^*_{ik}(\boldsymbol{\theta})$ for the probability $P(Z_i \geq \gamma_{i,k-1} | \theta_1, \ldots, \theta_n)$. Then the probability of a response in category $k$ can be written as

$$P_{ik}(\boldsymbol{\theta}) = P^*_{ik}(\boldsymbol{\theta}) - P^*_{i,k+1}(\boldsymbol{\theta}), \tag{8.12}$$

a generalization of the graded response model, which was introduced for one-dimensional latent space.

The next step is to define the distribution of errors $E_{ip}$. In this section we assume that the errors are normally distributed with variance $\sigma_i^2$. It follows that

$$P^*_{ik}(\theta) = \Phi\left(\frac{\sum_{j=1}^{n} \alpha_{ij}\theta_j - \gamma_{i,k-1}}{\sigma_i}\right). \tag{8.13}$$

For a one-dimensional model the probability $P^*_{ik}(\boldsymbol{\theta})$ is a normal ogive curve. The relation between (8.13) and the normal ogive model for dichotomous data (8.4) is discussed in exhibit 8.1.

Now, let the latent variables as well as the item scores be standardized; the thresholds are assumed to be defined in this metric. If we assume that the latent variables $\theta$ have a multivariate normal distribution, the response processes $Z$ are multivariate normally distributed. The marginal distribution of $Z_i$ is the standard normal distribution. Then

$$P(Z_i \geq \gamma_{i,k-1}) = \Phi(-\gamma_{i,k-1}).$$

Also, the joint distribution of any two variables $Z_i$ and $Z_j$ has the shape of a bivariate normal distribution. If the variables $Z$ had been observed, we would have used linear factor analysis to obtain the item parameters $\alpha$. Now we have the joint frequencies for pairs of variables $X_i$ and $X_j$,.

The marginal frequencies of the variables $X_i$ and the joint frequencies for pairs of variables can be used for the estimation of the factor loadings and the thresholds (Muthén, 1984). Of course, also the maximum likelihood approach, which is discussed in more detail in this chapter, can be used for the estimation of the item parameters. The maximum likelihood approach (Muraki & Carlson ,1995) uses the information of all answer patterns, and therefore is called full-information factor analysis.

The model does not have a pseudo-guessing parameter. In case guessing plays a role, some adaptations of the procedure are necessary. Bock et al. (1988) apply full-information factor analysis to dichotomous data. They discuss the effect of guessing and present analyses of a section of the LSAT uncorrected and corrected for guessing.

**Exhibit 8.1. The relationship between model (8.13) and the normal ogive model for dichotomous data and one-dimensional latent space**

What is the relationship between the model discussed in this section and model (8.4) for a one-dimensional latent space and dichotomously scored items?

With dichotomous data we can write the threshold model with normally distributed errors as

$$Z_{ip} = \theta_p + E_{ip},$$

$$\begin{aligned} X_{ip} &= 0 \quad \text{if} \quad Z_{ip} < b_i \\ X_{ip} &= 1 \quad \text{if} \quad Z_{ip} \geq b_i \end{aligned},$$

and

$$P_i(\theta) = \Phi\left(\frac{\theta - b_i}{\sigma_i}\right) = \Phi[a_i(\theta - b_i)],$$

where

$$a_i = 1/\sigma_i.$$

Without loss of generality we can assume that θ has a mean equal to zero and a variance equal to one. The variance of $Z_i$ is equal to

$$\sigma^2_{Z_i} = \sigma^2_\theta + \sigma^2_i = 1 + a_i^{-2}.$$

The correlation between $Z_i$ and θ is

$$\rho_i = \rho_{\theta Z_i} = \frac{\sigma^2_\theta}{\sigma_\theta\sqrt{\sigma^2_\theta + \sigma^2_i}} = \frac{a_i}{\sqrt{1 + a_i^2}}.$$

Let us now assume that θ is normally distributed. The correlation $\rho_i$ can be approximated by the biserial correlation between the item and the total test or rest-test, if the test is long. The biserial correlation estimates the correlation of a normally distributed variable assumed to underlie the dichotomous scores on a variable, with a continuous variable. The proportion correct of item *i* is equal to

$$\pi_i = \Phi\left(\frac{-b_i}{\sqrt{1 + a_i^{-2}}}\right).$$

It is clear that the item parameters $a_i$ and $b_i$ can be obtained from $\rho_i$ and $\pi_i$.

## 8.4 Item-Test Regression and Item Response Models

Consider Figure 8.6. It depicts the item-test regressions for two items of a 10-item test. The regressions do not correspond to fixed item characteristics. We would have obtained different regressions with different total tests and different examinee groups.

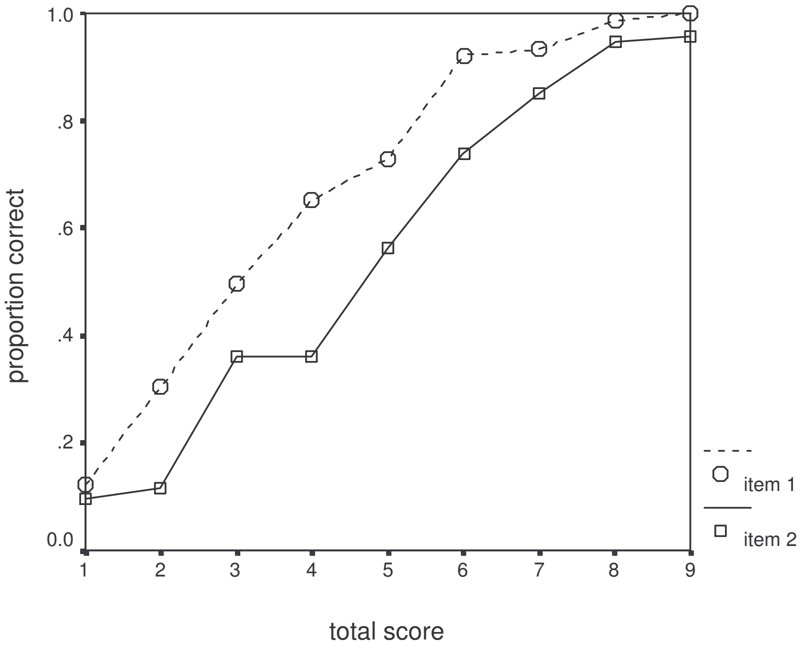


**FIGURE 8.6.** Item-test regression for two items from a 10-item test

Let us now look at Figure 8.7. We have the same items as in Figure 8.6. However, the abscissa is redefined. We have, for example, enlarged the difference between ‘8 correct’ and ‘9 correct’ in comparison to the difference between ‘5 correct’ and ‘6 correct’ (the first interval is nearly a factor 1.7 larger than the second interval). We have chosen different units and relabeled the axis as the latent trait dimension θ. In addition, we have fitted curves to the empirical regressions. The curves fit adequately. The curves are identical apart from a translation along the horizontal axis.

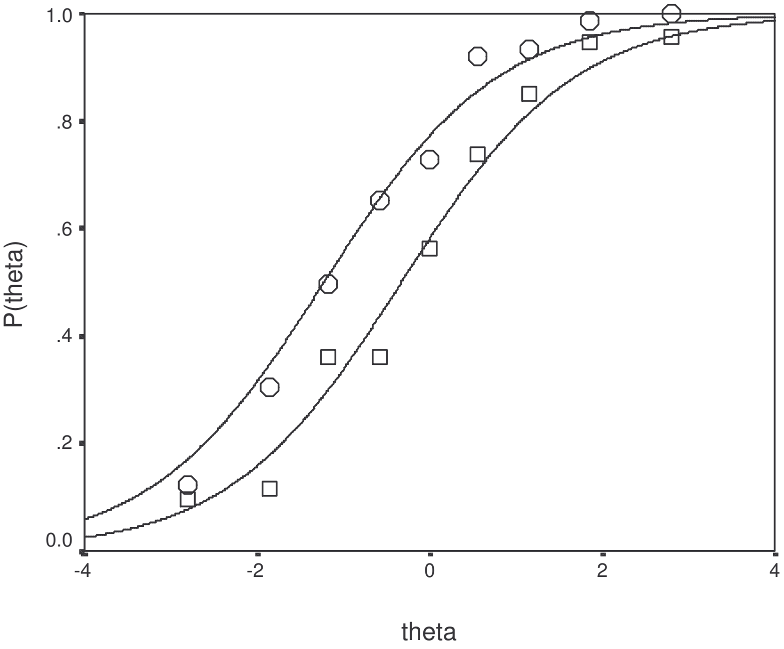


**FIGURE 8.7.** Estimated item-trait regressions (ICCs) for the two items from Figure 8.6

How did we proceed? We fitted the Rasch model to the data from the 10-item test. We used an estimation method in which the transformed total score is used as a proxy to θ. Is it correct to state that fitting the Rasch model or any other IRT model boils down to fitting item-test regressions?

Fitting an IRT model is not just fitting curves to item-test regressions for the following reasons:

- The item parameters are thought to be invariant item characteristics for the relevant population. The same two curves should obtain with different tests and examinee groups.
- Item parameters can be estimated even in case no examinee takes the same test. Also, in applications we do not have to think in terms of fixed tests. In CAT examinees respond to different items.

We mention some other relevant points:

- The estimation method used does not minimize a sum of weighted squared differences between the empirical regression and the curve. But such a weighted sum is used in one approach to determine item fit (see Section 8.14).
- The total score is used only for estimation in the Rasch model.
- Determining item fit is just one aspect of model fit. We must assess, for example, whether the data are unidimensional. Within a given score group the responses to two items should be independent (to be more precise, we expect a small negative correlation between the responses due to the fact that the item scores within a score group sum to a constant value).
- The estimation method used (JML, see Section 8.7) is statistically faulty. Examinees with the same total score do not have the same value for θ. As a result of the incorrect simplification in JML the item parameter estimates are biased and the ICCs in the figure are farther apart than they should be.

## 8.5 Estimation of Item Parameters

In a model with one latent trait the response probabilities given the latent trait are locally independent, as defined in (6.16). In case the responses are known, we call (6.16) the likelihood of the given response pattern instead of the probability. The likelihood of the responses $x_{ip}$ of $N$ persons ($p = 1,. . . , N$) on $n$ dichotomous items ($i = 1, . . . , n$) is given by

$$L = \prod_{p=1}^{N}\left(\prod_{i=1}^{n} P(X_{ip} = x_{ip} \mid \theta_p)\right) = \prod_{p=1}^{N}\prod_{i=1}^{n} P_i(\theta_p)^{x_{ip}}[1 - P_i(\theta_p)]^{1-x_{ip}}, \tag{8.14}$$

where $x_{ip} = 1$ for a correct response and $x_{ip} = 0$ for an incorrect response.

The main method for the estimation of the item parameters is the maximum likelihood method. There are three alternative maximum likelihood methods:

- Joint maximum likelihood, JML for short. In JML person and item parameters are estimated jointly.
- Marginal maximum likelihood (MML). In MML the person parameters are eliminated from the estimation process by integration over the distribution of person parameters.

- Conditional maximum likelihood (CML). In CML the person parameters are eliminated from the estimation process by conditioning on the total scores. The method is available for the logistic model with only a difficulty parameter. So, CML is possible with the Rasch model. Andersen (1983; see also Verhelst & Glas, 1995) demonstrated that CML is also possible when in the two-parameter model the slopes are assumed to be known.

Each of these methods will be discussed in the sequel: JML in Sections 8.6 and 8.7, MML in Section 8.8, and CML in Section 8.9.

Cohen (1979; see also Wright & Stone, 1979) proposed a simple approximation for the estimation of parameters in the Rasch model, assuming normally distributed item and person parameters. His approximation is based on the similarity between the logistic model and the cumulative normal distribution function. Urry (1974) suggested to estimate item parameters in the three-parameter normal ogive model from item indices under the assumption of normally distributed person parameters (see also Lord & Novick, 1968). The approximations might produce good starting estimates of parameters for the likelihood procedures.

For the multivariate normal distribution of latent abilities not only MML has been proposed, but also the analysis of marginals for items and item pairs (see Section 8.3).

## 8.6 Joint Maximum Likelihood Estimation for Item and Person Parameters

In the joint maximum likelihood (JML) method for dichotomous data person and item parameter estimates are obtained that maximize likelihood (8.14). First, starting values for the parameters are computed. Then, person parameters are computed that maximize the likelihood given the item parameter estimates. Next, new item parameter estimates are obtained on basis of the current estimates of the person parameters. One cycles through this process until the estimates from both sets of parameters are stable, i.e. until the changes in the estimates fall below a threshold.

Actually the natural logarithm of the likelihood, the log likelihood is maximized. Maximizing the log likelihood is easier than maximizing (8.14) itself and it produces the same estimates. Before the process is started persons with 0 responses correct and persons with perfect scores are eliminated; the maximum likelihood estimate of $\theta$ is $-\infty$ for a total score equal to 0 and $\infty$ for a total score equal to the number of items $n$. On similar grounds items that are answered correctly by all persons or by nobody, are removed.

The estimation procedure makes use of the fact that at its maximum a function has a zero slope. So, new parameter estimates are obtained by taking derivatives of the log likelihood with respect to these parameters, setting the results equal to zero and solving for the parameters (see Section 8.15). The estimation equation for person parameter $\theta_p$, the equation from which the new estimate of $\theta_p$ is to be obtained, can be written as

$$\sum_{i=1}^{n} \frac{x_{ip} - P_i(\theta_p)}{P_i(\theta_p)[1 - P_i(\theta_p)]} \frac{\partial P_i(\theta_p)}{\partial \theta_p} = 0\,, \tag{8.15}$$

where $\partial P_i(\theta_p)/\partial\theta_p$ is the derivative of $P_i(\theta_p)$ with respect to $\theta_p$.

The estimation equation for an item parameter of item $i$ is

$$\sum_{p=1}^{N} \frac{x_{ip} - P_i(\theta_p)}{P_i(\theta_p)[1 - P_i(\theta_p)]} \frac{\partial P_i(\theta_p)}{\partial \gamma_i} = 0\,, \tag{8.16}$$

where $\gamma_i$ is the item parameter of item $i$ in question.

There are $N$ estimation equations for person parameters and $n$ (Rasch model), $2n$ (2PL model) or $3n$ (3PL model) equations for item parameters. There is also one scale restriction (Rasch model) or two scale restrictions (2PL and 3PL model).

The estimation of parameters in the 2PL model and the 3PL model is not free of problems. In the 3PL model there are response patterns for which no unique maximum for θ exists (Samejima, 1973). The estimation of the lower asymptote $c_i$ might also give problems. In the 2PL model and the 3PL model the $a_i$-parameter estimates might be unstable. For that reason the change in parameter estimates from one iteration to the other is restricted in estimation programs. One way of restraining change is the introduction of prior distributions for the parameters (Swaminathan & Gifford, 1986).

## 8.7 Joint Maximum Likelihood Estimation and the Rasch Model

The Rasch model is the simplest model and for this reason very suitable for the introduction of JML. For the Rasch model the sets of equations for person parameters (8.15) and item parameters (8.16) can be simplified to

$$t_p = \sum_{i=1}^{n} P_i(\theta_p)\,,\; p = 1, \ldots, N, \tag{8.17}$$

where $t_p$ is the total score of person $p$, and

$$s_i = \sum_{p=1}^{N} P_i(\theta_p)\,,\; i = 1, \ldots, n, \tag{8.18}$$

where $s_i$ is the total number of correct responses to item $i$. In (8.17) the total score of a person is set equal to its expected value; in (8.18) the item total score is set equal to its expectation. The Equations (8.17) and (8.18) have to be solved iteratively for the parameters. One restriction must be added in order to fix the additive scale of the Rasch model.

From (8.17) it follows that all persons with the same total score have the same estimated θ. The total score is a sufficient statistic, a result that is valid only for the Rasch model. The implication is that JML estimation for the Rasch model can be viewed as a logistic regression problem with total scores and items as levels of two categorical variables.

In JML estimation the item parameter estimates depend on the person parameter estimates and vice versa. This causes an estimation problem of biased parameter estimates. The bias in the item parameters does not disappear if the number of persons $N$ increases. With large tests the bias is negligible (see Exhibit 8.2).

### Exhibit 8.2. The Failure of JML

Suppose we have one latent trait value, with $\theta = 0.0$, and two Rasch-items, with $b_1 = 0.0$ and $b_2 = 0.5$. The probability of a correct response to the items is:

$$p_1 = P_1(\theta = 0) = 0.5$$

and

$$p_2 = P_2(\theta = 0) = 0.37754.$$

We obtain the following probabilities:

$$\begin{aligned}
&P(1 \text{ correct; } 2 \text{ incorrect}) = p_1(1 - p_2) = \ .31123\\
&P(1 \text{ incorrect, } 2 \text{ correct}) = p_2(1 - p_1) = \underline{\ .18877}\\
&P(t(\text{total score}) = 1) \qquad\qquad\qquad = \ .5\\
&P(1 \text{ correct}|\ t = 1) = P(1 \text{ correct; } 2 \text{ incorrect})/\ P(t = 1) = \ .62246\\
&P(2 \text{ correct}|\ t = 1) = P(1 \text{ incorrect; } 2 \text{ correct})/\ P(t = 1) = \ .37754.
\end{aligned}$$

In JML we estimate the item parameters from the probabilities correct given the estimates of abilities. With two items we have only one ability estimate, for $t = 1$; this ability estimate can be set equal to 0.0 (which fixes the latent scale). So, we solve $b_1$ from $P(1 \text{ correct}|\ t = 1)$, and $b_2$ from $P(2 \text{ correct}|\ t = 1)$.

$$\hat{b}_1 = -\ln\left(\frac{P(1\,\text{correct}\mid t = 1)}{1 - P(1\,\text{correct}\mid t = 1)}\right) = -0.5\ ,$$

and

$$\hat{b}_2 = -\ln\left(\frac{P(2\,\text{correct}\mid t = 1)}{1 - P(2\,\text{correct}\mid t = 1)}\right) = 0.5\ .$$

The difference between the estimated $b$-parameters is twice the true difference. The problem with the JML-method is that it incorrectly equates the empirical regressions $P(i \text{ correct}|t)$ with $P_i(\theta)$.

Andersen (1972) demonstrated that with two Rasch-items the difference between the JML-estimates is always twice the true value. For a test with $n > 2$ Rasch items a correction factor $(n - 1)/n$ seems adequate for all practical purposes (Wright, 1988).

## 8.8 Marginal Maximum Likelihood Estimation

The marginal maximum likelihood method and the conditional maximum likelihood method do not exhibit bias in the item parameter estimates. In both methods person parameters are not estimated. In marginal maximum likelihood, or MML, we assume the existence of a distribution of $\theta$. We set the mean of this distribution equal to 0.0. This is allowed due to the interval property of the models (the additive scale property of the Rasch model). We also set the standard deviation of the latent distribution equal to 1.0 (this is permitted for the 2PL and the 3PL models). When we are interested in the Rasch model, we first estimate parameters of the 1PL model, i.e. the 2PL model with a common discrimination parameter. Next we transform the parameter estimates from the 1PL model to the Rasch model with a common discrimination parameter equal to 1.0.

Of course we do not know the population distribution of person parameters. We have to make some assumptions with respect to the distributional form. The choice we make has an effect on the values of the item parameter estimates. This dependency seems to be the weak spot in MML. Fortunately, it is possible to estimate properties of the latent distribution along with the item parameters when enough data are available.

In case we do not have any information on the distribution of person parameters, a natural choice is a more or less bell-shaped distribution. Traditionally the obvious choice is the normal distribution. A disadvantage is, however, that the distribution leads to awkward computations. This problem can be overcome: we can approximate the normal distribution by a discrete distribution with any degree of accuracy.

Let us consider a discrete distribution of θ with $q$ latent classes. The relative frequencies of $\theta_k$ ($k$ = 1, . . .,$q$) are denoted by $g(\theta_k)$. When the values $\theta_k$ and $g(\theta_k)$ are chosen for an optimal approximation of a continuous distribution, they are called quadrature points and quadrature weights, respectively.

The probability of response pattern $\mathbf{x} = (x_1, . . ., x_i, . . .,x_n)$ given $\theta_k$ is written as $P(\mathbf{x}|\theta_k)$. If a randomly chosen person from the population makes the $n$-item test, the probability of response pattern **x** equals

$$P(\mathbf{x}) = \sum_{k=1}^{q} P(\mathbf{x} \mid \theta_k) g(\theta_k)\,. \tag{8.19}$$

Now we administer the test. $S$ response patterns occur. Response pattern $l$ ($l$ = 1 , . . .,$S$) occurs $r_l$ times.

In MML we maximize the natural logarithm of

$$L_{\text{MML}} = C \prod_{l=1}^{S} P(\mathbf{x}_l)^{r_l}\ , \tag{8.20}$$

where $C$ is independent of the parameters (Bock & Aitkin, 1981). The resulting estimation equations can be written as

$$\sum_{k=1}^{q} \frac{n_{ik} - n_k P_i(\theta_k)}{P_i(\theta_k)[1 - P_i(\theta_k)]} \frac{\partial P_i(\theta_k)}{\partial \gamma_i} = 0\,, \tag{8.21}$$

where

$\gamma_i$ is one of the item parameters of item $i$,

$n_{ik}$ = the posterior expectation of the number correct on the item in latent class $k$,

and

$n_k$ = the posterior ‘size’ of latent class $k$.

In the EM-algorithm we solve iteratively for maximum likelihood estimates and update values $n_{ik}$ and $n_k$ in the so-called ‘Expectation’ step until convergence is reached.

## 8.9 Conditional Maximum Likelihood Estimation in the Rasch Model

In the Rasch model we also can apply the conditional maximum likelihood (CML). It turns out that in discussing CML another model representation is preferable. We rewrite the Rasch model as:

$$P_i(\theta) = \frac{\exp(\theta - b_i)}{1+\exp(\theta - b_i)} = \frac{\xi\varepsilon_i}{1+\xi\varepsilon_i} \tag{8.22}$$

where $\xi$=exp($\theta$) and $\varepsilon_i$ = exp(-$b_i$).

Assume that we have two items with item parameters $\varepsilon_1$ and $\varepsilon_2$. The probability that item 1 is correctly answered, given one correct response on the 2-item test is equal to the probability of the score pattern (1,0), divided by the sum of the probabilities of the score patterns (1,0) and (0,1). The probability is

$$P(x_1 = 1 \mid x_1 + x_2 = 1, \xi) = \frac{\dfrac{\xi\varepsilon_1}{1+\xi\varepsilon_1} \times \dfrac{1}{1+\xi\varepsilon_2}}{\dfrac{\xi\varepsilon_1}{1+\xi\varepsilon_1} \times \dfrac{1}{1+\xi\varepsilon_2} + \dfrac{1}{1+\xi\varepsilon_1} \times \dfrac{\xi\varepsilon_2}{1+\xi\varepsilon_2}} = \frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2}. \tag{8.23}$$

The probability in (8.23) is independent of the value of the person parameter $\xi$. The comparison between items can be made independent of the value of the person parameters (and vice versa the comparison between person parameters can be made independent of the items). The measurements are *specific objective*.

With three items we can do the same thing as with two items. The probability that item 1 is correct given a total score $t$ = 1 on the three-item test is

$$P(x_1{=}1, x_2{=}0, x_3{=}0 | t{=}1) = \varepsilon_1/(\varepsilon_1+\varepsilon_2+\varepsilon_3).$$

The probability that item 1 is correct given a total score equal to 2 is:

$$P(x_1{=}1,x_2{=}1,x_3{=}0|t{=}2) + P(x_1{=}1,x_2{=}0,x_3{=}1|t{=}2) = (\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3)/(\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3).$$

This result can be generalized to a test with $n$ Rasch-items. First, some notation is introduced. In the denominator of $P(x_i = 1 \mid t = 2)$ all combinations $\varepsilon_j\varepsilon_{j'}$ appear; in the numerator only the combinations with the parameter of item $i$, $\varepsilon_i$, are entered. The sum of the products is called an elementary symmetric function. The elementary symmetric function for four items and a total score equal to 2, the elementary symmetric function of order 2, is

$$\gamma_2(\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4) = \gamma_2(\boldsymbol{\varepsilon}) = \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_1\varepsilon_4 + \varepsilon_2\varepsilon_3 + \varepsilon_2\varepsilon_4 + \varepsilon_3\varepsilon_4\,.$$

The denominator of $P(x_i = 1 \mid t = 2)$ can be written as $\gamma_2(\boldsymbol{\varepsilon})$. For the numerator of $P(x_i = 1|t = 2)$ $\gamma_1^{(1)}(\boldsymbol{\varepsilon})$, the elementary symmetric function of order 1 exclusive item 1, is needed. The numerator can be written as $\varepsilon_1\gamma_1^{(1)}(\boldsymbol{\varepsilon})$. The elementary symmetric function of order 0 is defined to be equal to one.

In the general case there are $n$ items. The number of correct responses $s_i$ to item $i$ ($i$ = 1, . . ., $n$) is based on the group of $N$ persons with $0 < t < n$ responses correct. The conditional likelihood is equal to

$$L_{\text{CML}} = \text{Prob}(s_1,...,s_n \mid t_1,...,t_N) = \frac{\text{Prob}(s_1,...,s_n,t_1,...,t_N)}{\text{Prob}(t_1,...,t_N)}$$

$$= C\frac{\prod_{i=1}^{n}\varepsilon_i^{s_i}}{\prod_{p=1}^{N}\gamma_{t_p}(\boldsymbol{\varepsilon})} \tag{8.24}$$

where $C$ is a factor independent of the item and person parameters. The estimation equation for the estimation of $\varepsilon_i$ is

$$s_i = \sum_{t=1}^{n-1} N_t P(x_i = 1 \mid t) = \sum_{t=1}^{n-1} N_t \frac{\varepsilon_i \gamma_{t-1}^{(i)}(\boldsymbol{\varepsilon})}{\gamma_t(\boldsymbol{\varepsilon})}, \tag{8.25}$$

where $N_t$ denotes the number of persons with score $t$ ($t$ = 1,. . . $n$ - 1). So, in CML the item score $s_i$ is set equal to the expected item score based on the conditional probabilities correct for the various total scores. The estimation equations (8.25) for the item parameters $\varepsilon_i$ ($i$ = 1, . . ., $n$) are solved iteratively under one constraint needed in order to fix the latent scale.

## 8.10 More on the Estimation of Item Parameters

Special attention must be given to the presence of missing values when item parameters are estimated. There are two cases to consider: data can be missing by design or not. Data are missing by design when e.g. the number of items is too large to present all items to an examinee. Then subtests can be administered to different examinee groups. The subtests must have common items in order to obtain item and person parameter estimates on a common scale (see Chapter 10). All three maximum likelihood methods can be generalized to the incorporate values missing by design. For the MML-approach the consequence is that latent distributions for multiple groups must be defined. Data also can be missing because items are skipped. In achievement testing skipped items sometimes are treated as incorrect responses. Another, more adequate approach to deal with skipped items with the multiple-choice format is suggested by Lord (1980). When the presence of missing values correlates with latent ability data is not missing at random (Little & Rubin, 1987). MML-estimation of item parameters is affected. A special case of missing values arises when the test is speeded, i.e. when some examinees do not reach the items at the end of the test. Clearly, in that case blind application of a IRT model is inadequate.

For accurate estimation of a difficulty parameter it is important that the group of persons that took the test, has an average ability level comparable to the item difficulty. In the 2PL model and the 3PL model also discrimination parameters must be estimated. These parameters define the slopes of the ICCs. Information on the steepness of a slope is available only when the latent abilities are reasonably well spread. The Rasch model does not have a discrimination parameter. In the Rasch model item parameter estimation can be accurate even in case all persons have the same ability. This advantage is of limited value, however: if all abilities are equal, there is no way of discriminating between alternative models. The estimation of $c$ in the 3PL model is more accurate when we have more relatively low abilities.

Inaccurate estimation of the pseudo-chance-level parameter has an impact on the estimation of the discrimination parameter and the difficulty parameter as well, for the estimates of the item parameters are correlated. For known abilities the inverse of the matrix of error variances and covariances of the item parameter estimates, the information matrix of the item parameters, is given in Lord (1980).

The CML estimation procedure for the Rasch model has a clear statistical advantage above the other estimation procedures as has been discussed in the previous section. For the Rasch model software has been developed for the estimation of item parameters with CML; CML-estimation also can be done with a special kind of log-linear analysis (Heinen, 1996; Kelderman, 1984). CML is, however, computationally demanding. This problem might be avoided by using MML in which the characteristics of the population distribution are estimated along with the item parameters (De Leeuw & Verhelst, 1999). There is another disadvantage of using software for the Rasch model. If the Rasch model does not fit the data very well, other models should be fitted and other software is needed. The Rasch-model can be viewed as a submodel of the 2PL model and the 3PL model. There is much to say for using the same software to compute item parameter estimates for alternative models and to compare the outcomes of these models.

Software for the analysis of item responses is commercially available, for well-known item response models like the IRT models for dichotomous data discussed here as well as for other models. Information on software can be found in books and articles that describe applications or research with the software, from software houses and from software review sections of journals like *Applied Psychological Measurement*. Embretson & Reise (2000), who introduce many of IRT models, discuss a selection of the commercially available computer programs:

> TESTFACT (Wilson, Wood & Gibbons, 1991) for the full-information factor analysis of dichotomous data with the two- and three-parameter normal ogive models (with fixed values *c*),
> BILOG (Mislevy & Bock, 1990) for the estimation of the 1PL, 2PL and 3PL models, and BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) for the analysis of multiple groups (BILOG is no longer available; see BILOG-MG3 of Assessment Systems Corporation),
> MULTILOG (Thissen, 1991) and PARSCALE (Muraki & Bock, 1993) for dichotomous as well as polytomous items,
> XCALIBRE (Assessment Systems Corporation, 1996) for the estimation of parameters in the 2PL and 3PL model,
> and RUMM (Sheridan, Andrich & Luo, 1996) for the estimation of parameters of various Rasch models.

The authors notice the fact that no final review of software is possible while programs have been revised and will be revised continually. They also notice that alternative programs may unexpectedly produce different results although model specifications are identical. So, more comparative studies on IRT programs and possible flaws of certain programs have to be done. It is to be hoped that this leads to improvements of the IRT software.

### 8.11 Maximum Likelihood Estimation of Person Parameters

Assume that the item parameters have been estimated – with CML or MML – and the person parameters have not been estimated yet. Then the person parameters can be estimated by maximum likelihood given the estimated item parameters, i.e. in a way similar to estimation in JML. In this section we will discuss the estimation of person parameters by ML (for a bias in ML, see Warm, 1989).

The person parameter in each of the models can be obtained by solving Equation (8.15) for θ. We rewrite this equation as

$$\sum_{i=1}^{n} w_i(\theta)x_i = \sum_{i=1}^{n} w_i(\theta)P_i(\theta)\,, \tag{8.26}$$

with

$$w_i(\theta) = \frac{P_i'(\theta)}{P_i(\theta)[1-P_i(\theta)]}\,, \tag{8.27}$$

where $P_i'(\theta) = \partial P_i(\theta)/\partial\theta$ .

In the equation a weighted total score is set equal to the weighted sum of the expected item scores. The weight in (8.27) can be compared to optimal weight (4.8) for congeneric measurements.

In the 3PL model the size of the weight $w_i(\theta)$ depends on the value of θ. For $c_i > 0$ the weight decreases as θ becomes smaller; at θ = -∞ the weight is zero. In the 3PL model one must verify whether the solution to (8.26) is a maximum. Equation (8.26) always is solved by setting θ equal to -∞, but only in some aberrant cases there is a maximum of the likelihood at θ = -∞. In some cases with aberrant response patterns two maxima exist, one for θ = -∞.

For the 2PL model optimal weight (8.27) is equal to the discrimination parameter $a_i$. In the Rasch model, in which the discrimination parameters are equal, all weights are equal to 1. In the Rasch model the left hand side of (8.26) reduces to the total score; the total score is sufficient for the estimation of θ (see Equation (8.17)).

Several important properties of IRT models presuppose the ML approach. The use of the information function as a measure of precision presupposes maximum likelihood estimation of θ. In Computerized Adaptive Testing either the maximum likelihood estimate of θ or a Bayesian estimate, the subject of the next section, is used.

In some applications, such as the scoring of a group of examinees on one and the same test, there is no obligation to use maximum likelihood. Then the question arises whether to use optimal scoring weights or not. For the Rasch model this question is easily answered. In this model the maximum likelihood estimator is a nonlinear transformation of the unweighted total score on the test. So, total score can be computed, and reported either on the observed score scale or on a transformation of this scale. Things are different with respect to the 3PL model and the 2PL model.

With the 2PL model the optimal item weights are equal to the item discrimination parameters $a_i$. One might wonder how much gain in accuracy is obtained by using these weights instead of the unweighted total score. This is of some importance, for the unweighted total score has the advantage that the scoring rule needs not much explaining to the examinees. There is a second reason to be careful with using the optimal weights. The optimality of the weights is based on the assumption that the item parameters are very accurately estimated and that the IRT-model fits the data.

It is easy to imagine that in many situations the gain from using optimal weights is only apparent. If the discrimination parameters $a_i$ differ moderately the gain in accuracy when using weighted scores, is limited. An additional problem can be inaccuracy of the item parameter estimates. Using ‘optimal’ weights based on inaccurate parameter estimates might result in less accurate estimates of abilities than unit weighting. Then it is advisable to use unit weights just for statistical reasons.

In the 3-PL model there is a further complication in connection with the differential weighting of items: the optimal item weight does not only depend on the item parameters, but also on the unknown person parameter. Fortunately in some applications optimal weights can be chosen that do not depend on ability level. This is the case when differentiation between ability levels is really needed only near a particular ability level, $\theta_0$. In that case weights $w_i(\theta_0)$ are appropriate for all values of $\theta$ (Lord, 1980, p. 170) and the problem of weighted scores is equal to the problem of weighting in the 2PL model.

Tests have been scored with unit item-weights even though methods for weighting items have been proposed many times, frequently for good reasons. The developments in IRT will not end this practice.

An IRT analysis remains very useful even when it has been decided to use unit weights for the combination of the item scores. In Chapter 6 it has already been demonstrated that the outcome of an IRT analysis can be used for the computation of the conditional standard error of measurement.

## 8.12 Bayesian Estimation of Person Parameters

In MML we work with a distribution of $\theta$, with characteristics fixed beforehand or estimated from the test data. Our knowledge of the distribution can be used to obtain a Bayesian estimate of $\theta$.

With a discrete distribution of $\theta$ we can write the posterior probability distribution for $\theta$ given the responses to the items as

$$P(\theta_k \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \theta_k)g(\theta_k)}{P(\mathbf{x})} = \frac{P(\mathbf{x} \mid \theta_k)g(\theta_k)}{\sum_{h=1}^{q} P(\mathbf{x} \mid \theta_h)g(\theta_h)}, \; k = 1, \ldots, q \qquad (8.28)$$

where $g(\theta_k)$ is the prior probability of $\theta_k$ and $P(\mathbf{x}|\theta) = L(\mathbf{x}|\theta)$ is the likelihood of the observed scores $\mathbf{x} = (x_1, \ldots, x_n)$ given $\theta$.

There are two alternative Bayesian estimators of $\theta$, the posterior mean or EAP estimator (*Expected A Posteriori estimator*) and the posterior mode. With a discrete distribution the EAP estimator is the obvious choice. The estimator is

$$\mathrm{EAP}(\theta) = \sum_{k=1}^{q} \theta_k P(\theta_k \mid \mathbf{x})\,. \qquad (8.29)$$

The posterior mean is very easy to compute: no iterations are needed in order to obtain it. The posterior variance can be computed as a measure of uncertainty.

## 8.13 Test and Item Information

In classical test theory the variance of measurement errors was a relevant concept. The inverse of the error variance is an indicator of the precision with which statements about persons can be made. In IRT it appears advantageous to start with the precision of measurements. The two central concepts are *test information* and *item information*.

The test information at a level of latent ability θ, $I(\theta)$, gives – under some conditions - the precision with which θ can be estimated at this ability level. The conditions are:

a) We have chosen the adequate model;
b) The item parameters are known, at least accurately estimated;
c) We use maximum likelihood estimation, i.e. we use optimal weights for the estimation of θ; with other estimation methods (e.g. with number right scoring) we speak of the information of a scoring formula; the latter information cannot exceed the test information.
d) The test is not too short.

The test information has a very convenient property. It is the sum of the item informations (Birnbaum, 1968)

$$I(\theta) = \sum_i I_i(\theta) . \qquad (8.30)$$

So, the contribution of each item to the accuracy of a test may be considered apart from the contributions of other items.

The item information of item *i* can be written as

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)[1 - P_i(\theta)]} , \qquad (8.31)$$

where $P'_i(\theta)$ is the derivative of $P_i(\theta)$ with respect to θ. So, the item information is equal to the square of the slope of the ICC at θ divided by the local error variance, which is the item variance given θ. For polytomous items results are given in Exhibit 8.3.

In the 3PL model the item information is

$$I_i(\theta) = a_i^2 p_i(\theta)[1 - p_i(\theta)]\left(1 - \frac{c_i}{P_i(\theta)}\right), \qquad (8.32)$$

where $p_i(\theta)$ equals the probability of a correct response to the item if $c_i$ would have been equal to 0. In the 2PL model the item information equals $a_i^2$ times the item variance given θ, and in the Rasch model the item information equals the item variance given θ, the error variance on the true-score scale.

### Exhibit 8.3. Optimal weights and Information of Polytomous Items

Optimal weight (8.27) and the item information (8.31) cannot directly be generalized to the case of polytomous items. With polytomous items each option has an optimal option weight. Let $P_{ik}(\theta)$ be the option characteristic curve. Then the optimal weight associated with option *k* is

$$w_{ik}(\theta) = \frac{\partial \ln P_{ik}(\theta)}{\partial\theta} = \frac{P'_{ik}(\theta)}{P_{ik}(\theta)} .$$

The sum of the weights of the chosen options equals zero at the maximum likelihood estimate. The option weights function in quite a different way than the weights for dichotomous items. The reason for the difference is that the weight for a dichotomous item is an item weight: the options *correct* and *incorrect* are not weighted separately, as the score for *incorrect* is set equal to zero. The item weight is the difference between the option weight for *correct* and the option weight for *incorrect*.

For the two models (8.6) and (8.7) the score weight of category *k* can be written as *k* plus a factor independent of *k*. The option score *k* can be written as

$$k = w_{ik}(\theta) - w_{i0}(\theta) .$$

While *k* does not depend on the item and option parameters, the polytomous Rasch models have a sufficient statistic for the estimation of θ: the sum of the category numbers of the options chosen.

The item information, the sum of the option informations, is

$$I_i(\theta) = \sum_{k=0}^{m} \frac{P'^2_{ik}(\theta)}{P_{ik}(\theta)} .$$

It is easily demonstrated that the item information of a dichotomous item, as given by (8.31), is a special case of the item information for polytomous items.

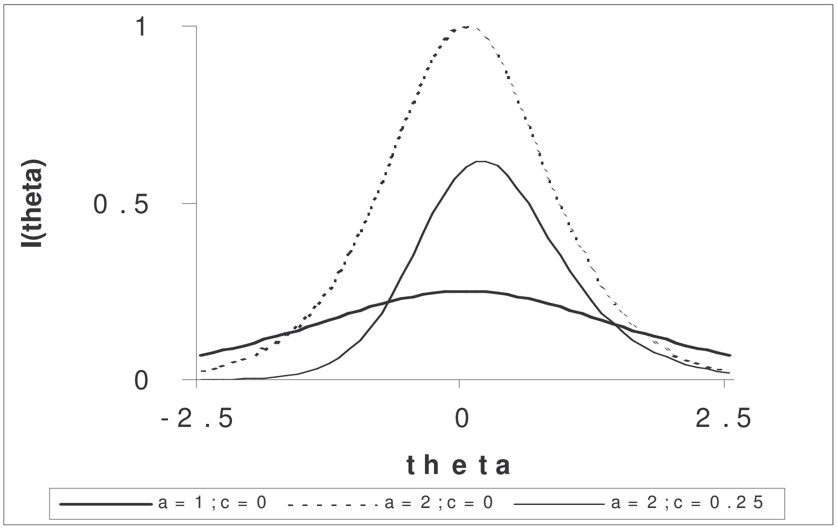


**FIGURE 8.8.** Information functions for 3 items ($b = 0$)

In Figure 8.8 information functions are displayed for three items. The information of the item with $c$ = 0.0 and $a$ = 2.0 exceeds the information of the item with the lower discrimination parameter for a large range of the latent ability. With increasing $a$ the information at $\theta = b$ increases, while the information at more distant abilities decreases. A perfect Guttman item discriminates only at one point: it discriminates between persons with θ smaller than $b$ and persons with θ larger than $b$.

The information of the item with $c$ = 0.25 and $a$ = 2.0 is lower than the information of the item with the same value of $a$ and $c$ = 0.0. The reduction is lowest at high values of θ; for low values of θ the reduction is large due to guessing. From the figure we can infer that the highest information is obtained at $\theta = b$, unless $c$ is larger than 0. When $c$ exceeds 0, the highest information value is obtained for a value of θ somewhat higher than $b$. Birnbaum (1968) gives the relationship between $c$ and the value of θ at which the highest information is obtained.

The value of the item information, and consequently the value of the test information, depends on the choice of the latent scale. In the 2PL model and 3PL model a linear transformation of the scale is allowed. We can multiply all θ and $b$ with 2 and divide all $a$ by 2. Then the item information and test information decrease by a factor 4 (see (8.32)). And, when also nonlinear transformations of the latent scale are considered - for example, a transformation to the true-score scale - the form of the information function can change dramatically.

So, information is not an invariant item property. The ratio of the item informations of two items is, however, invariant:

$$\frac{I_i(\theta^*)}{I_j(\theta^*)} = \frac{I_i(\theta)}{I_j(\theta)}$$

for all monotone transformations $\theta^*$ of θ.

Also the *relative efficiency* of two tests, the ratio of the test informations of two tests, remains unchanged with a change of scale. This means that the comparison of the accuracy of two tests does not depend on the chosen latent scale.

The estimated value of $I(\theta)$ can be used (asymptotically) for the construction of a confidence interval for θ. The variance of $\hat{\theta}$ equals the inverse of the test information, $1/I(\theta)$, assuming accurate item parameter estimates (otherwise the error variance is larger, see De Gruijter, 1988). Under the assumption that $\hat{\theta}$ is normally distributed, the approximate 95-percent confidence interval is

$$\hat{\theta} - 1.96/\sqrt{I(\hat{\theta})} < \theta < \hat{\theta} + 1.96/\sqrt{I(\hat{\theta})}\,. \qquad (8.33)$$

With confidence interval (8.33) we might err in case a population of abilities is involved. Then we better use the EAP estimator instead of the maximum likelihood estimator. With this estimator we can also compute the posterior variance of θ. This variance is smaller than the inverse of the test information. These results are comparable to those discussed in connection with the application of the Kelley formula within the context of classical test theory.

## 8.14 Model-data fit

An important question is whether the chosen IRT model fits the data. If the model does not fit, we have to find a less restrictive model that does fit the data or we have to drop nonfitting items. The investigation of model fit has two aspects

a a statistical test of model fit;
b an analysis of residuals, in order to get an idea of the seriousness of the model violations.

With a small number of data it is difficult to reject a particular model. With a large number of data a model is easily rejected even if the deviations of the data from the model are small for all practical purposes. For this reason both statistical testing and the analysis of residuals are necessary in the study of model fit.

In the models we have discussed so far we have one general model assumption: the assumption of unidimensionality or local independence. For each model we also have assumptions regarding the specific shape of the item response curves. The assumption of unidimensionality can be verified in several ways:

a A factor analysis might reveal whether the data are unidimensional. A factor analysis of phi coefficients - ordinary product moment correlations - is not suitable when the item discriminations are high. The analysis of phi coefficients would produce spurious factors. A multidimensional IRT-analysis, which is a nonlinear factor analysis, is called for. A factor analysis of tetrachoric correlations - these are correlations of bivariate normal distributions assumed to underlie the responses to pairs of dichotomous items - might be appropriate when the latent distribution is normal and there is no guessing.

b We can examine the item variances and covariances after the elimination of the model effects. A positive correlation between residuals indicates violation of model assumptions. Stout (1987) presented a nonparametric test for unidimensionality based on a split of the test into two subtests (see also Roussos, Stout & Marden, 1998).

Specific model assumptions can be verified in various ways. Varying values of item-test correlations might indicate different discrimination parameters. The study of item-test regressions can reveal whether guessing is likely to play a role. We also can do an analysis of residuals by item, after having completed an IRT analysis. First we group the estimated θ's in a number of sufficiently large groups. Next in each group the (standardized) difference between the observed proportion correct $p$ and the model probability is computed. Yen's (1981) statistic $Q_1$ is based on the squared differences. The statistic she proposed is a Pearson chi-square. For item $i$ the statistic can be written as

$$Q_{1i} = \sum_{l=1}^{m} \frac{N_l [p_{il} - P_i(\theta_l)]^2}{P_i(\theta_l)[1 - P_i(\theta_l)]} = \sum_{l=1}^{m} z_{il}^2 \ , \tag{8.34}$$

where $P_i(\theta_l)$ stands for the average probability correct in ability group $l$ and $N_l$ is the size of ability group $l$. $Q_{i1}$ is approximately distributed as a chi-square with $m - k$ degrees of freedom, where $m$ equals the number of score groups and $k$ is the number of item parameters. A likelihood ratio statistic has been proposed by Mislevy & Bock (1990)

$$G_i^2 = -2\ln\frac{L_{i1}}{L_{i0}} = 2\sum_{l=1}^{m}\left( N_{il} \ln\frac{p_{il}}{P_i(\theta_l)} + (N_l - N_{il})\ln\frac{1-p_{il}}{1-P_i(\theta_l)} \right), \tag{8.35}$$

where $L_1$ is the likelihood given the model, $L_0$ the likelihood without model restrictions and $N_{il}$ the number of correct responses to item $i$ in group $l$. This statistic is also approximately chi square distributed. A problem with the $Q_1$ and $G$ statistics is that the grouping of persons is based on estimated values θ. In a simulation study on new tests of fit and the $G$ statistic from (8.35) by Glas and Falcón (Glas and Falcón, 2003) it was demonstrated that this $G$ statistic has an inflated type I error rate.

Figure 8.9 shows the fit of an item of the verbal analogies test, a subtest of a Dutch intelligence test, the NDT or Netherlands Differentiation Test (NDT, 2002, in prep.) using the likelihood ratio fit test (8.35). The figure includes the item response curve and the empirical regression of the item.

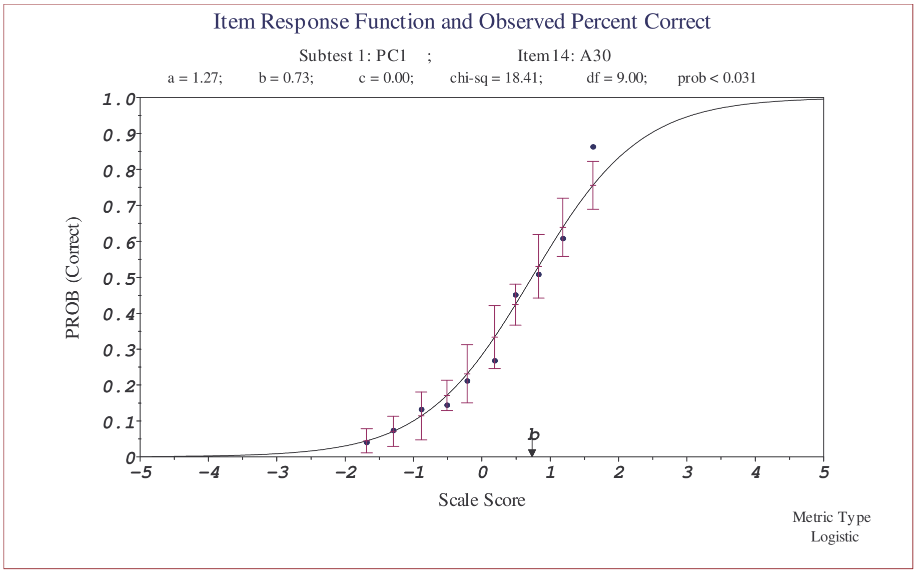


**FIGURE 8.9**. Item fit for an item of a Verbal Analogies Test of the NDT, the Netherlands Differentiation Test (BILOG output)

Other chi-square statistics – on the item and the test level – have been proposed for the Rasch model, based on the CML approach (Andersen, 1973; Kelderman , 1984; Molenaar, 1983).

In MML an overall test to compare nested models seems to be possible (Reise, Widaman & Pugh, 1993).

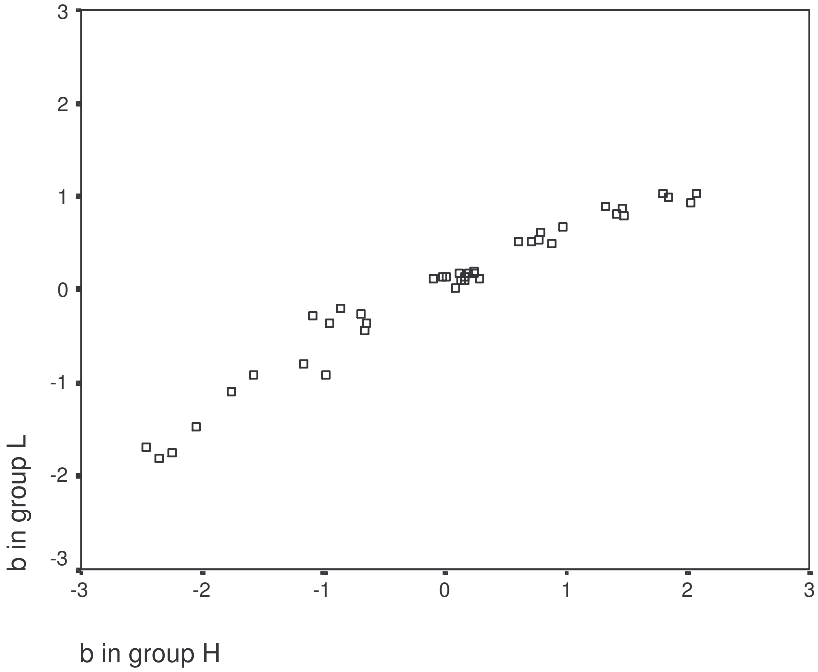


**FIGURE 8.10.** Rasch item parameter estimates in a low scoring group and a high scoring group

The residuals used for the computation of item fit statistics might also be plotted. Graphical model control might add useful information on the cause of item misfit. Also for other checks on model fit graphical methods are useful. An example is given in Figure 8.10 where Rasch item parameters have been estimated in a group with high scores (H) and a group with low scores (L). The item parameter estimates are not invariant. In the low scoring group the range of values of the item parameter estimates is smaller. This is due to the fact that guessing played a role in the (simulated) data. A good overview of graphical methods is given by Hambleton, Swaminathan & Rogers (1991).

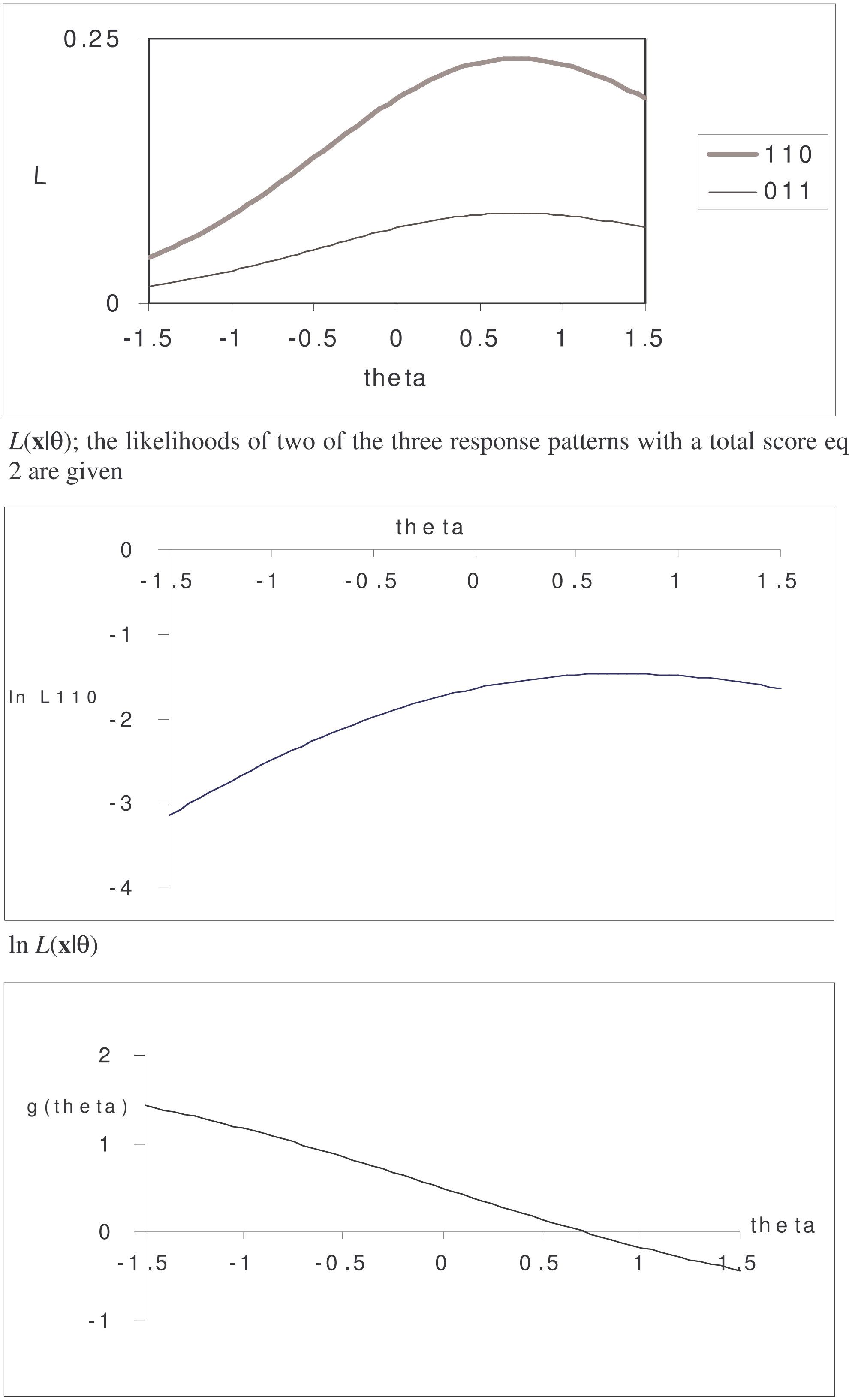
## 8.15 Appendix: Maximum Likelihood Estimation of θ in the Rasch Model

In this appendix we give an example of the way maximum likelihood parameter estimation proceeds. For simplicity we have chosen the estimation of the person parameter in the Rasch model.

The likelihood of a score pattern $\mathbf{x} = (x_1, \ldots, x_n)$ in the Rasch model can be written as

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \frac{\exp(\theta - b_i)^{x_i}}{1 + \exp(\theta - b_i)} = \exp(\theta)^t \prod_{i=1}^{n} \exp(-b_i)^{x_i} \prod_{i=1}^{n} [1 + \exp(\theta - b_i)]^{-1}, \qquad (8.36)$$

where $t$ is the total score. We want to find the value of θ that maximizes (8.36). Maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood. So, instead of maximizing (8.36) we maximize its logarithm



a) $L(\mathbf{x}|\theta)$; the likelihoods of two of the three response patterns with a total score equal to 2 are given

b) $\ln L(\mathbf{x}|\theta)$

c) $g(\theta)$

**FIGURE 8.11.** The likelihood $L(\mathbf{x}|\theta)$, $\ln L(\mathbf{x}|\theta)$ and the derivative of $\ln L(\mathbf{x}|\theta)$ with respect to $\theta$ for a test with $b_1 =$ -0.5; $b_2 = 0.0$; $b_3 = 0.5$, and a total score equal to 2

$$\ln L(\mathbf{x} \mid \theta) = t\theta - \sum_{i=1}^{n} x_i b_i - \sum_{i=1}^{n} \ln[1 + \exp(\theta - b_i)]\,, \tag{8.37}$$

where ln is the natural logarithm. When ln $L(\mathbf{x}|\theta)$ has obtained its maximum as a function of θ the derivative of ln $L(\mathbf{x}|\theta)$ with respect to θ is equal to 0. So we can find the ML estimate of θ by differentiating (8.37) with respect to θ and setting the result equal to 0 (we must check whether we have obtained a maximum of the function and not a minimum).

Differentiating (8.37) with respect to θ and setting the result equal to 0 gives the equation

$$g(\theta) = t - \sum_{i=1}^{n} \frac{\exp(\theta - b_i)}{1 + \exp(\theta - b_i)} = t - \sum_{i=1}^{n} P_i(\theta) = 0\,. \tag{8.38}$$

This equation is identical to (8.17).

In Figure 8.11 the likelihood, the logarithm of the likelihood and the derivative of ln $L(\mathbf{x}|\theta)$ are displayed for a simple example. In Figure 8.11.a we can see that two response patterns with identical total scores have very different values for the likelihood, but also that for both response patterns the maximum is obtained for the same value of θ (0.721). In Figure 8.11b ln $L(\mathbf{x}|\theta)$ is given. From Figures 8.11a and 8.11b it is clear that the value of θ that maximizes $L(\mathbf{x}|\theta)$ also maximizes ln $L(\mathbf{x}|\theta)$. Also for this value of θ $g(\theta)$ in Figure 8.11.c equals 0.0. In Figure 8.11c we see that the slope of $g(\theta)$ is negative. This shows that we have obtained a maximum instead of a minimum of ln $L(\mathbf{x}|\theta)$.

We find θ by solving (8.38). We have a maximum in (8.38) if $g(\theta)$ decreases with increasing θ in the neighborhood of $g(\theta) = 0.0$. In other words, the derivative of $g(\theta)$ with respect to θ, the second order derivative of ln $L$, must be negative. The derivative of $g(\theta)$, $g'(\theta)$, can be used in the estimation procedure. This is done in the Newton-Raphson procedure.

Let us demonstrate the estimation procedure. We approximate the function $g(\theta)$ at the value of the estimated θ in iteration $k$, $\theta^k$, by a straight line. This line goes through the point $(\theta^k, g(\theta^k))$ and has a slope equal to $g'(\theta^k)$. The equation of this line is

$$g(\theta) = g(\theta^k) + g'(\theta^k)\,(\theta - \theta^k). \tag{8.39}$$

Setting $g(\theta)$ equal to 0 gives us the next estimate of θ:

$$\theta^{k+1} = \theta^k - g(\theta^k)/\, g'(\theta^k)\,. \tag{8.40}$$

In the Newton-Raphson method we need $g'(\theta)$, the derivative of $g(\theta)$ in (8.38):

$$g'(\theta) = -\sum_{i=1}^{n} P_i(\theta)[1 - P_i(\theta)]\,. \tag{8.41}$$

The derivative of $g(\theta)$ is equal to minus the test information. This relation between the second order derivative of the log likelihood, $g'(\theta)$, and the test information does not hold for the 2PL model and the 3PL model. With these models the test information equals minus the *expected* value of $g'(\theta)$ (Kendall & Stuart, 1961, p.8-9). In the 2PL model and the 3PL model the test information is easier to compute than the second order derivative $g'(\theta)$, so with these models we replace the second derivative in (8.39) by minus the test information.

The iterative procedure for the estimation of θ in the Rasch model is:

A We compute a starting value for θ, $\theta^0$.
B We compute a new value $\theta^1$ by application of (8.40):

$$\theta^1 = \theta^0 - \frac{\left(t - \sum_{i=1}^{n} P_i(\theta^0)\right)}{\left(-\sum_{i=1}^{n} P_i(\theta^0)[1 - P_i(\theta^0)]\right)}. \tag{8.42}$$

C We compute $|\theta^1 - \theta^0|$, the absolute value of the difference between the two consecutive estimates of θ.

D If the value obtained in the C-step is below a chosen threshold value ε we stop: the obtained value $\theta^1$ is our final ML estimate. If the difference exceeds the threshold we replace $\theta^0$ by $\theta^1$ and repeat the steps B and C. This process is repeated until we have reached convergence (by the way: the method may fail to converge in some maximization problems).

Let us give a numerical example of the method. We have three items, with $b_1$ =-0.5, $b_2$ = 0.0; and $b_3$ = 0.5: The total score $t$ equals 2. As starting value we choose θ = 0.0. The final estimate of θ is .721, obtained in the second iteration. The data of the iteration process are given in the table below. You might want to verify these figures yourself using a spreadsheet.

| iteration | $\theta^k$ | $g(\theta^k)$ | $g'(\theta^k)$ | $\theta^{k+1}$ |
|---|---|---|---|---|
| k = 0 | 0.0 | 0.5 | -0.72001 | 0.69444 |
| k = 1 | 0.69444 | 0.01706 | -0.64820 | 0.72075 |
| k = 2 | 0.72075 | 0.00007 | -0.64304 | 0.72086 |

## Exercises

8.1 Compute the probability of a correct response for a Rasch item with item parameter equal to 0.0 and person parameter θ = -2.0 (0.5) 2.0.

8.2 We have the responses of two homogeneous groups of persons on two items. The response probabilities are: $P_1(\theta_1)$ = 0.3775, $P_1(\theta_2)$ = 0.6225, $P_2(\theta_1)$ = 0.4378 and $P_2(\theta_2)$ = 0.7112. Estimate the person parameters $\theta_1$ and $\theta_2$ on basis of the response probabilities for the first item, assuming that the Rasch model fits. Use the response probabilities for the second item for verifying whether the Rasch model really fits.

8.3 Given is a test with three Rasch items. The item parameters are: $b_1$ = -0.5, $b_2$ = 0.0 and $b_3$ = 0.5. A person has answered items 1 and 2 correctly, and item 3 incorrectly. Compute the likelihood for θ = -1.0, -0.5, 0.0, 0.5, 1.0.
a) Consider the four intervals defined by the five values of θ for which the likelihood has been computed. In which interval lies the maximum likelihood estimate of θ?

b) Assume that we have a population distribution with five latent classes: $\theta_1 = -1.0$, $\theta_2 = -0.5$, $\theta_3 = 0.0$, $\theta_4 = 0.5$, $\theta_5 = 1.0$. Also assume that these latent classes have the same relative frequencies: $g(\theta_k) = 0.2$ for $k = 1, \ldots, 5$. Compute the EAP estimate of θ.

8.4 We have three items with item parameters:

$b_1 = 0.5$, $a_1 = 1.0$, $c_1 = 0.0$
$b_2 = 0.5$, $a_2 = 2.0$, $c_2 = 0.0$
$b_3 = 0.5$, $a_3 = 2.0$, $c_3 = 0.25$.

Compute the item informations at $\theta = 0.0$.

8.5 We have a discrete distribution of θ with values –1, -0.5, 0.0, 0.5 and 1. The following is known:

| value θ | frequency $f(\theta)$ | $I(\theta)$ |
|---|---|---|
| -1.0 | 0.1 | 7.864 |
| -0.5 | 0.2 | 9.400 |
| 0.0 | 0.4 | 10.000 |
| 0.5 | 0.2 | 9.400 |
| 1.0 | 0.1 | 7.864 |

Compute the reliability of the test when maximum likelihood is used for the estimation of θ.

# 9 APPLICATIONS OF IRT

## 9.1 Introduction

IRT can be used to analyze a test and to perform an item analysis. In section 9.2 item analysis with IRT will be discussed. There is more to IRT. The development of IRT has opened new ways for test applications and research with tests. This is true especially for the parametric IRT models. Nonparametric models may be less restrictive than parametric models, but they are also less informative.

An important application of IRT is test equating, or bringing test scores to the same scale. IRT equating has greatly extended the possibilities of equating. Large-scale testing programs also profit from developments in IRT. Different persons may get different, partially overlapping tests. Test results from all persons and all items can be brought to the same scale using IRT. We will devote the next chapter to equating with and without IRT.

Item banking is an important tool with IRT-based testing. Assume that a new test is administered that contains old items, items with known item parameters, and also a number of new items. After the test administration we can estimate the item parameters of the new items on the common latent scale as well. By repeatedly applying this procedure we are building a large pool of items with known item parameters. As long as the item characteristics do not change due to, for example, educational change (the phenomenon of item drift; Donoghue & Isham, 1998) we have an *item bank* from which we may choose items with known item parameters at will.

One application of item banks is with test construction. In the application of standard test-development techniques of classical test theory to the construction of tests, items are selected on the basis of two statistical characteristics: item difficulty and item discrimination. What IRT has to offer for test development in general, and item selection in particular, is described in Section 9.3.

IRT enables us also to investigate item bias (see 9.4) and inappropriate response patterns (9.5) in a better way than classical test theory.

Another important application of IRT is in so-called computerized adaptive testing. Because under ML scoring items can be viewed as the building blocks of tests, an item bank enables to administer computerized adaptive tests (CAT), the subject of Section 9.6.

The use of IRT in the measurement of change is discussed in section 9.7. Finally, IRT makes it possible to tackle various problems by doing simulation studies. An example, discussed in Exhibit 9.1, is the determination of the optimal number of options in multiple choice items. In Section 9.8 some concluding remarks on IRT-applications are made.

**Exhibit 9.1. On the Optimal Number of Options in Multiple-Choice Items**

Items with four answer options are more accurate than items with three or two options, ceteris paribus: the effect of guessing is smaller with these items. On the other hand, more two-option items can be answered in the same testing time than four-option items. So, it is a sensible question to investigate whether one should administer a test with, say, two-option items rather than a test with four-option items. There have been studies on this topic from both a theoretical as well as an empirical perspective. Let us follow Lord's (1977) line of reasoning.

First, we must decide how many items with a particular number of options can be administered within a given testing time. Lord assumed that the reading time depends on the number of options alone, i.e., the number of items times the number of options is fixed for a given testing time. This assumption is testable for a particular application.

Lord also assumed that item characteristics do not change with a change of the number of choices, except the value of the pseudo-chance-level parameter. The pseudo-chance-level parameter $c$ is set equal to the inverse of the number of options although this decision is not supported by the outcome of IRT-research.

Lord showed that 3-option items provide the most information at the midrange of the scale score, whereas the 2-option item works best at the upper range. When pass-fail decisions must be made, tests with two or three options are optimal given optimal item difficulties and the validity of the assumptions mentioned above.

The bulk of the research on the optimal-number-of-options problem is done from an empirical perspective. The most recent study, it seems, is done by Rogers and Harley (1999). Their overall conclusion is that tests consisting of 3-option items are at least equivalent to tests composed of 4-option items in terms of internal consistency. Haladyna (1999) strongly recommends 3-option items instead of 4- or 5-option items. His recommendations, however, are to no avail at all. The bulk of test constructors and item writers stick to 4- or 5-options with multiple-choice items.

## 9.2 Item Analysis and Test Construction

IRT can be used to investigate the dimensionality of a test and to screen the items of the test. Some remarks with respect to item analysis have been made in Chapter 6. In this section we proceed by giving two examples of IRT-analyses of tests.

IRT-modeling is frequently used with achievement testing. Our first example is about the IRT-analysis of measurement instrument from another domain. The instrument is a personality questionnaire, the Multidimensional Personality Questionnaire (Tellegen, 1982), and the IRT-analysis was done by Reise & Waller (1990).

Reise and Waller did several analyses of the scales of the MPQ. They did, for example, a factor analysis on the tetrachoric correlations. They concluded that the responses on each scale could be accounted for by one dominant dimension. The responses of two samples of 1000 persons were analyzed with the one- and two-parameter logistic models. The overall fit of the two-parameter model was adequate although some items did not fit well. Reise and Waller concluded that the IRT analysis gave more information on the psychometric properties of the scales and the items than would have been possible with an analysis based on the classical test model.

The second example is about the application of IRT in the construction of a measurement instrument. Van der Rijt, Van Luit & Pennings (1999) describe the construction of two versions or scales of the Utrecht Early Mathematical Competence Scales. The scales were developed in order to assess the developmental level of early mathematical competence in children ages 4 to 7 years. First, items were written for eight aspects of numerical and nonnumerical knowledge of quantity. This resulted in a pool of 120 items, from which two forty-item scales had to be constructed.

The total test of 120 was too large to be presented to the children participating in the investigation. So, several test booklets were constructed. Common items would make it possible to obtain item parameters on a common scale. First, item difficulties were computed and a factor analysis was done on the scores of the eight aspects. The results of the factor analysis suggested that there is one underlying dimension of mathematical competence.

The data were analyzed with the Rasch model. This model was rejected. Next, an analysis was done with the two-parameter model in which the values of the discrimination parameters *a* were fixed. With fixed discrimination parameters the item difficulty parameters could be obtained with a CML-analysis (Verhelst & Glas, 1995). Some items did not fit well. Among the non-fitting items were items that were frequently guessed correctly. The non-fitting items were eliminated from the item pool and the two forty-item scales were composed of items from the pool.

## 9.3 Test Construction and Test Development

In this section it is assumed that a large set of items is available for test construction and that accurate item parameter estimates of these items have been obtained. This could have been achieved by an analysis of the responses of examinees to different test forms with common items.

When good estimates of the item parameters are available, we can compute the item information. The item information has the additivity property under ML-scoring of persons: the sum of the item informations produces the test information. This entails that we can compute test characteristics beforehand from the characteristics of the items that we choose for the test. So, we can construct the shortest test that at a certain ability level has a test information exceeding a specified minimum value. To simplify matters we assume that there are no additional restrictions on the test composition.

Suppose that we want to construct a test that has an error variance of maximally the value $d$ at ability level $\theta_0$. Then the test information at that point, $I(\theta_0)$, should have at least the value $1/d$. We construct the shortest test as follows. We take as first item in the test the item with the highest value of the item information at $\theta_0$. As second item we choose the item that has the highest item information among the remaining items. We go on with selecting items until the sum of the item informations at $\theta_0$ is at least equal to $1/d$. Unfortunately parameter estimates are not free from error. So, when selecting on item information we might err a bit and the computed test information might be somewhat higher than the true test information. The procedure is simple. The largest problem is the determination of the value $d$. When an old test is available, we could use the value $d$ of that test.

Another possibility is that we want to specify a minimum value of the test information for two (cf. Exhibit 6.2) or more values of θ. With this conceptually simple extension the solution becomes quit difficult.

Let us take the case where we want to discriminate for a range of θ values, and where a minimum value of θ must be specified for an interval of θ. In this case we replace the interval by some well-chosen values of θ. Next, for each of these values θ a minimum value for the test information is specified. Now, we need to find the shortest test for which at each of these values θ the test information is at least as large as the minimum specified. This problem can be defined as an optimization problem, to be solved with linear or integer programming techniques (Theunissen,1985; Van der Linden & Boekkooi-Timminga, 1989).

We formulate the problem as follows. Determine the minimum level for the information at $\theta_k$, $I(\theta_k)$ ($k$=1, . . ., $m$). Minimize test length

$$n = \sum_{i=1}^{N} x_i \ , \tag{9.1}$$

where $N$ is the number of items in the item pool and $x_i$ =1 if the item is included in the test and 0 otherwise, subject to the constraints

$$\sum_{i}^{N} x_i I_i(\theta_k) \geq I(\theta_k) \qquad (k = 1, \ldots, m) \qquad (9.2)$$

Also in this case the specification of the target value of the test information might appear the most difficult part of the problem. In Chapter 6 we referred to an investigation by Cronbach & Warrington (1952), one of the first IRT-based simulation studies. They demonstrated that in most cases the highest average test accuracy is obtained with items of a similar difficulty, at the cost of a loss in accuracy at the higher and lower abilities. This is demonstrated in Figure 9.1. In this figure the test informations of two five-item tests with moderately discriminating items (all $a$ equal to 1) are contrasted. One test is a peaked test, a test in which all $b$-parameters are equal. In the second test the difficulty parameters are spread, with $b$ = -1.0, -0.5, 0.0, 0.5 and 1.0. The first test is optimal in a large interval around $b$ = 0. The study by Cronbach and Warrington suggests that we might set the target information lower for the relatively high and low abilities.

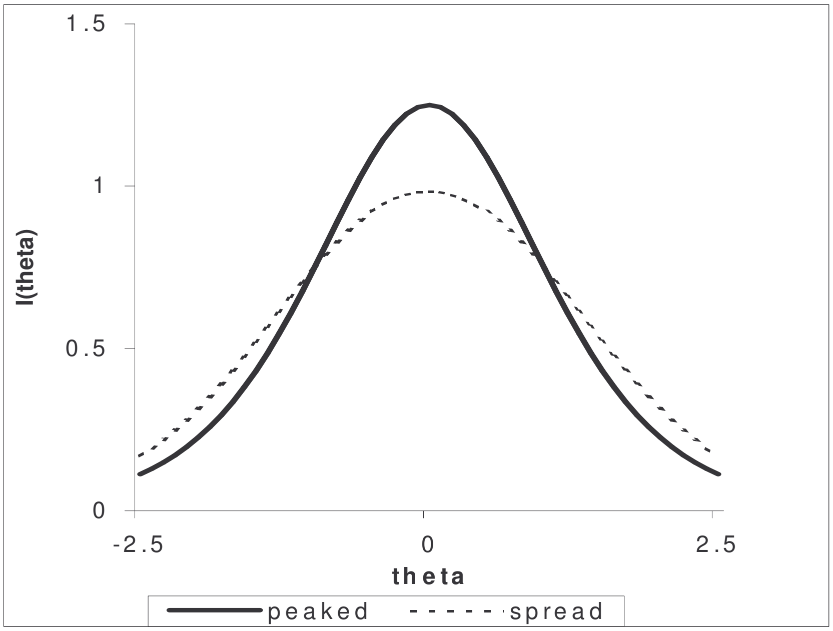


**FIGURE 9.1.** The test information of a peaked and a spread test

An interesting example of the possibilities of optimal test construction is presented by Van der Linden and Reese (1998). They demonstrate the possibilities of test construction with extra constraints on, for example, subject matter coverage in the context of computerized adaptive testing. *Applied Psychological Measurement* has dedicated a special issue to optimization problems (Van der Linden, 1998).

## 9.4 Item Bias or DIF

A test and the use of test scores are meant for a well-defined population or several distinct populations. Test takers from a population generally can be classified into several subgroups, e.g. *male* and *female*, and inferences from the test scores should be equally valid for members of all subgroups. Bias is defined as *differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers* by Cole & Moss (1989, p. 205)

Bias can be studied in various ways. A possibility that frequently is considered, is that a test is potentially biased because of the presence of biased items. So, bias can be studied by investigating the internal structure of a test.

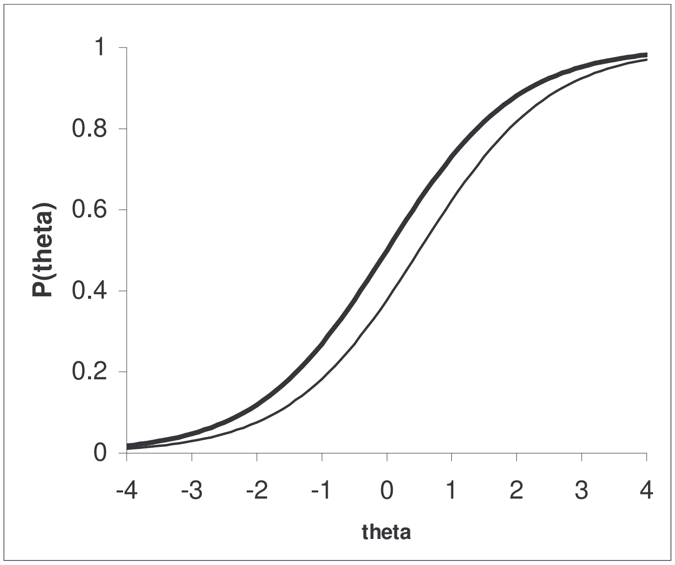
In terms of IRT we speak of item bias if the probability of a correct response to the item *given latent ability* differs between subgroups. Figures 9.2.a and 9.2b illustrate item bias. In Figure 9.2.a we have uniform item bias. One subgroup has a disadvantage at each level of θ; the ICC for this subgroup lies below the ICC for the other subgroup at each level of θ. Figure 9.2.b illustrates the second possibility of item bias. In this figure the ICCs cross. At the left of their intersection one subgroup has the advantage, at the right the other subgroup has the advantage.

To date, the term differential item functioning (DIF) is used for item bias. In the 1999 *Standards* (APA, AERA, & NCME, 1999) DIF is defined as a statistical property of a test item in which different groups of test takers who have the same total test score have different average item scores, or in some cases, different rates of choosing various item options (l.c. p. 175). Technically, and in IRT specifically, DIF is a special kind of violation of the unidimensionality assumption. A single latent trait, i.e. the value of θ, is not sufficient to predict the probability of a correct response; besides latent ability group membership is necessary in order to correctly predict the probability. This is most easily seen when in the Rasch model the biased item is shifted *f* units to the right for one subgroup, say subgroup 2:

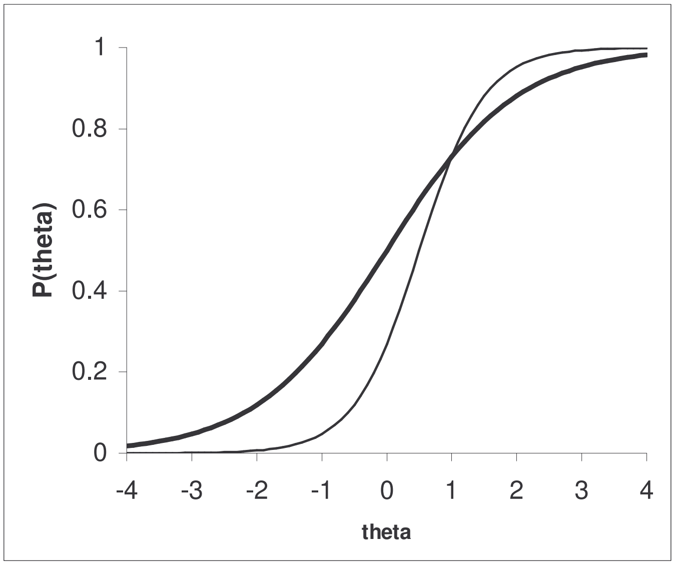
$$P_i(\theta_{p(g)}) = \frac{\exp(\theta_p - b_{i(g)})}{1+\exp(\theta_p - b_{i(g)})}, \text{ where } b_{i(2)} = b_{i(1)} + f.$$

An alternative way of writing the ICC's for both subgroups is:

$$P_i(\theta_{1p(g)}, \theta_{2p(g)}) = \frac{\exp(\theta_{1p} + \theta_{2(g)} - b_i)}{1+\exp(\theta_{1p} + \theta_{2(g)} - b_i)}, \text{ where } \theta_{2(1)} = 0.0,\ \theta_{2(2)} = -f.$$



(a) uniform DIF



(b) nonuniform DIF

**FIGURE 9.2.** Two examples of DIF; the ICCs of one item in two different subgroups

If DIF is observed, it is important to identify the causes of the effect. Is the item less familiar in the focal group, the group we are interested in, than in the reference group, and, if so, is the differential familiarity unrelated to the ability of interest? If the difference between the performance of two groups is really due to irrelevant factors, we can conclude that the item is biased, which, of course, is undesirable. A recent example of research into the causes of DIF is the study by Allalouf, Hambleton & Sireci (1999). These investigators addressed the causes of DIF in a test translated from Hebrew into Russian. Several sources of DIF were found: changes in word difficulty, changes in item format, differences in cultural relevance, and changes in context.

Many methods for the detection of DIF have been proposed. Let us review some of them in an IRT context. Rudner, Getson & Knight (1980) proposed to look at the size of the deviation between the ICCs in both subgroups over the relevant range of θ. In several proposals for measuring item bias the differences between the ICCs are counted only for values of θ found in the study (Shepard, Camilli, & Williams, 1985).

In their study on item bias Linn and Harnisch (1981) computed the item parameters on basis of the results of all subjects in the study. For each person they could compute the probability of a correct response given the estimate of ability. They compared the average model proportion correct for a range of values θ to the observed proportion correct in the target group. They also proposed to compute standardized differences. In this respect their approach is similar to Yen's approach to item fit. Linn and Harnisch did not propose to use a $\chi^2$ statistic, however.Thissen, Steinberg & Wainer (1988) proposed a likelihood ratio statistic for an item suspected to be biased. First, they proposed to estimate the item parameters for both groups simultaneously. The resulting likelihood for this model, $M_0$, is $L_0$. Next, the item parameters can be estimated again, but the item parameters of the item under investigation are allowed to take on different values in the two groups. The likelihood for this alternative model, $M_1$, is $L_1$. The likelihood ratio statistic equals

$$LR = -2\ln\frac{L_0}{L_1}\,. \tag{9.3}$$

*LR* is approximately $\chi^2$ distributed under the null hypothesis (equal item parameters in different groups). The degrees of freedom are equal to number of parameters set equal in both groups in model $M_0$, but are allowed to vary between groups in model $M_1$. More information on this approach is given by Kim & Cohen (1998), who use the *LR* test in connection with the graded response model.

Some methods for the detection of DIF are not based on the IRT approach. Two of these methods deserve our attention: *STD P-DIF* (See Exhibit 9.2), and the Mantel-Haenszel measure and chi-square statistic (See Exhibit 9.3). Both approaches are based on an analysis of the two-by-two tables of group membership and correct versus incorrect responses for the various total score levels. Both approaches can be related to the IRT-approach, with observed score as an indicator of latent ability. The Mantel-Haenszel method works well in case all items are Rasch items and the investigated item is the only biased item (Zwick, 1990). When the items are not Rasch-items and the ability distributions of the focal and reference groups differ, one better takes another procedure for the detection of DIF. The procedure SIBTEST (Shealy & Stout, 1993) corrects for the fact that the raw score indicates different expected true scores in the both groups.

When several items are biased the detection of DIF becomes more difficult: the biased items have too much influence on the latent ability estimate. Therefore Kok, Mellenbergh & Van der Flier (1985) proposed an iterative method for the detection of DIF. Millsap & Everson (1993), Camilli & Shepard (1994), and Scheuneman & Bleistein (1999) give overviews of methods for the detection of item bias. Cole & Moss (1989) make some critical remarks with respect to DIF methodology and the interpretation of the outcomes of DIF-studies.

Finally, an interesting approach related to DIF-research is the mixed model approach proposed by Rost (1990, 1991). In this approach a person does not belong to a group defined on basis of an external criterion (e.g. female, male), but to a latent class. In each latent class a unidimensional model is assumed to hold, but the item parameters differ between latent classes. The analysis might reveal the existence of classes of respondents who solve the test problems according to different strategies, resulting in different item difficulties. When there are two latent classes, a big one and a small one, the small latent class generates deviant response patterns, the subject of the next section.

### Exhibit 9.2. A simple index for DIF: *STD P-DIF*

We have administered a test and for one item computed the item-test regression, both for the reference group, group R, as for the focal group, group F. The latter group is the group of interest. We have obtained the following values for the item-test regressions (proportions correct given total score *k*) and score frequencies *n*:

| Total score | $p_{\mathrm{R}k}$ | $n_{\mathrm{R}k}$ | $p_{\mathrm{F}k}$ | $n_{\mathrm{F}k}$ |
|---|---|---|---|---|
| 0 | 0.0 | 0 | 0.0 | 0 |
| 1 | 0.3000 | 10 | 0.2500 | 4 |
| 2 | 0.4000 | 30 | 0.3333 | 3 |
| 3 | 0.4588 | 85 | 0.4286 | 7 |
| 4 | 0.4818 | 110 | 0.4667 | 15 |
| 5 | 0.5133 | 150 | 0.4444 | 9 |
| 6 | 0.7143 | 140 | 0.6667 | 12 |
| 7 | 0.8538 | 130 | 0.8125 | 16 |
| 8 | 0.8800 | 100 | 0.8182 | 22 |
| 9 | 0.9556 | 45 | 0.9167 | 12 |
| Total | 0.6575 | 800 | 0.6600 | 100 |

$$\textit{STD P-DIF} = \sum_k n_{\mathrm{F}k}\left(p_{\mathrm{F}k} - p_{\mathrm{R}k}\right) / \sum_k n_{\mathrm{F}k} = p_{\mathrm{F}} - \sum_k n_{\mathrm{F}k}\, p_{\mathrm{R}k} \,/\, n_{\mathrm{F}} = -0.0452$$

We see from the table that 4 persons from the focal group have a total score equal to 1 and that 1 of them (a proportion equal to 0.25) has the item correct. In the reference group the proportion correct given a total score equal to 1 is 0.3.

With *STD P-DIF* – shorthand for “standardized P-difference” - we compute the weighted mean difference between the proportions correct for the focal group and those for the reference group. As weights we use the proportions of persons from the focal group with the respective total scores. Index *STD P-DIF* has been proposed as a DIF index by Dorans & Kulick (1986; see also Dorans & Holland, 1993). *STD P-DIF* can be used to detect uniform bias. In the example above we have obtained the index for a very short test; this has been done in order to keep the computational burden small. The value of the index is -0.0452: the item is more difficult for the focal group. The proportions correct in the total groups are 0.6575 for the reference group and 0.66 for the focal group. So, in this example the focal group has a higher overall achievement on the item, but nevertheless the item seems to be biased against this group. Actually, the effect is small. Even with large groups only standardized differences larger than 0.05 and smaller than -0.05 are considered for further inspection.

Total score has been used in the computations for the index as an indicator of ability for members of both groups. Notice too that the item that possibly displays DIF is included in the computation of the total score.

Total score is sufficient for the estimation of ability under the assumptions of the Rasch model. More importantly, in the Rasch model the item proportion correct given total score does not depend on characteristics of the latent trait distribution.

**Exhibit 9.3. The Mantel-Haenszel procedure**

At score level $k$ the probability of a correct response to an item is $p_{1rk}$ in the reference group, and $p_{1fk}$ in the focal group, the group we are interested in. In the Rasch model the odds in the reference group, $p_{1rk}/(1 - p_{1rk}) = p_{1rk}/p_{0rk}$, is equal to the odds in the focal group, $p_{1fk}/p_{0fk}$. So, the odds ratio is equal to 1. The Mantel-Haenszel measure estimates the extent to which the odds ratio deviates from 1. The Mantel-Haenszel estimator for the item is

$$\alpha_{\mathrm{MH}} = \frac{\sum_{k=1}^{s} n_{1rk} n_{0fk} / n_k}{\sum_{k=1}^{s} n_{0rk} n_{1fk} / n_k},$$

where $n_{1rk}$ is the number of persons at score level $k$ belonging to the reference group and having a correct response to the item, etc. In practice a logarithmic rescaling of $\alpha$ is used as a measure of effect size.

The Mantel-Haenszel chi-square statistic (dropping a continuity correction) is

$$\chi^2_{\mathrm{MH}} = \frac{\left( \sum_{k=1}^{s} n_{1fk} - \sum_{k=1}^{s} \mathrm{E} n_{1fk} \right)^2}{\sum_{k=1}^{s} \mathrm{Var}(n_{1fk})},$$

where

$$\mathrm{E} n_{1fk} = \frac{n_{1k} n_{fk}}{n_k},$$

and

$$\mathrm{Var}(n_{1fk}) = \frac{n_{1k} n_{0k} n_{rk} n_{fk}}{n_k^2 (n_k - 1)}.$$

## 9.5 Deviant Answer Patterns

Not only items may show deviations from the model specifications. Also persons can have responses that deviate from the pattern expected in the IRT model. An example is a low scoring person who copies part of the answers of a high scoring neighboring examinee (Wollack & Cohen, 1998). Wollack and Cohen defined the following index for the similarity between the responses of an alleged Copier and a Source:

$$\omega = \frac{h_{CS} - \sum_{i=1}^{n} P(u_{iC} = u_{iS} \mid \theta_C)}{\sqrt{\sum_{i=1}^{n} P(u_{iC} = u_{iS} \mid \theta_C)[1 - P(u_{iC} = u_{iS} \mid \theta_C)]}}, \tag{9.4}$$

where

$n$ = number of items

$C$ = Copier
$S$ = Source
$u_{iC}$ = response of $C$ to item $i$
$u_{iS}$ = response of $S$ to item $i$
$h_{CS}$ = number of identical responses.

A high value of $\omega$ indicates that copying probably has been occurred.

Various general methods have been proposed for the detection of deviant answer patterns or the analysis of person fit as it is also known in the literature. Drasgow, Levine & McLaughlin (1987) discuss a number of indices, among them the standardized $L_0$, wich easily can be generalized to models for polytomous items (Drasgow, Levine & Williams, 1985), and the fit statistic $F2$.

$L_0$ is the logarithm of the likelihood evaluated at the maximum likelihood estimate of $\theta$. An atypical response pattern is indicated by a low log likelihood of the response pattern $u_1$ . . .$u_n$ given the estimated ability parameter. But, what is a relatively low log likelihood? For an appropriate interpretation of the log likelihood the expected value and the standard deviation of the log likelihood given the estimated person parameter are needed. The standardized index in connection with dichotomous items is

$$z_3 = \frac{L_0 - M(\hat{\theta})}{s(\hat{\theta})}, \tag{9.5}$$

with

$$L_0 = \sum_{i=1}^{n} \left[ u_i \ln P_i(\hat{\theta}) + (1 - u_i) \ln[(1 - P_i(\hat{\theta})] \right]. \tag{9.6}$$

The mean and variance of the log likelihood given the estimated value of $\theta$ are

$$M(\hat{\theta}) = \sum_{i=1}^{n} \left[ P_i(\hat{\theta}) \ln P_i(\hat{\theta}) + [1 - P_i(\hat{\theta})] \ln[(1 - P_i(\hat{\theta})] \right] \tag{9.7}$$

and

$$s^2(\hat{\theta}) = \sum_{i=1}^{n} P_i(\hat{\theta})[1 - P_i(\hat{\theta})]\{\ln P_i(\hat{\theta}) - \ln[1 - P_i(\hat{\theta})]\}^2. \tag{9.8}$$

Molenaar & Hoijtink (1990) noticed that $L_0$ does not have a normal distribution. They proposed to use a chi-square distribution for the evaluation of deviant answer patterns.

The $F2$ statistic was suggested by Rudner (1983) as a generalization of a fit statistic used in connection with the Rasch model. The statistic is

$$F2 = \frac{\sum_{i=1}^{n} [u_i - P_i(\hat{\theta})]^2}{\sum_{i=1}^{n} P_i(\hat{\theta})[1 - P_i(\hat{\theta})]}. \tag{9.9}$$

Drasgow et al. (1987) noticed that the statistics did not uniformly well. They suggested that it might be necessary to construct separate indices for different kinds of deviant response patterns (spuriously low and spuriously high patterns). Of course, also special purpose indices might be developed, like ω.

### 9.6 Computerized Adaptive Testing (CAT)

Tests can be computer administered. Nowadays a wide variety of item formats is available in computer-based tests, both items with a forced item response as well as items with open response formats. Computerized testing makes it possible to allow different examinees to take a test on different occasions. For each examinee a different test can be composed, in order to avoid the risk that items become known among other things. Tests frequently are composed using a stratified random selection procedure. In that case results can be analyzed with generalizability theory, and, when items are scored dichotomously, with approaches discussed in Chapter 6.

With computerized testing more is possible. It is possible, for example, to use a sequential testing design. One example of such an approach is the closed sequential procedure mentioned in Chapter 6.

With item response theory computerized testing can be made even more flexible. First, consider a traditional test. Such a test is meant for measurements in a population of persons, the target population. No test can be equally accurate for all persons from the target population. However, with computerized adaptive testing we have the possibility to administer each person a test in such a way that the test score is as accurate as possible.

If we knew the ability of a person we could administer a test tailored to the ability level of this person. However, we do not know the ability level of a person; if we knew there would be no need of testing. Using item response theory a testing strategy can be used such that step by step a person's ability is estimated. At each consecutive step the estimate is more precise. The choice of the item or subset of items at each step is tailored to the estimated ability at the previous step. This calls for items for which item parameters have already been estimated. All the items are stored in an item bank, and for this large set of items IRT item parameter estimates are known.

More technically, for the administration of the first item we can start with the not unreasonable assumption that a person to be tested is randomly chosen from the target population. The population distribution can be regarded as the prior distribution for the ability of this person. After each response we can compute the posterior distribution of $\theta$ from the prior distribution $g(\theta)$ and the likelihood of all responses $L(\mathbf{x}|\theta)$ (cf. (8.28)). This distribution can be used as the new prior distribution for the ability of the person. We choose a new item optimal with respect to this prior. We might, for example, after a response to an item compute the posterior mean, the EAP estimate, and select a new item that has the highest item information at the level of the EAP estimate. After a correct response the estimated ability is higher than after an incorrect response. Therefore a more difficult item is administered after a correct response than after an incorrect response. We might stop when the error variance is smaller than some criterion. When the EAP estimate is used, the relevant error variance is the posterior variance of $\theta$ (Bock & Mislevy, 1982). This CAT-procedure is illustrated in Figure 9.3. For practical reasons another stopping rul is frequently used: the test length of the CAT-procedure is fixed.

Sometimes it is profitable to redefine the unit of presentation in CAT and to group items into testlets. One argument for grouping could be that several items are based on the same subject or the same text, but there might be other reasons for grouping items as well (Wainer

& Kiely, 1987). With a redefinition of the unit of presentation a different choice of item response model might be in order (Wainer & Wang, 2000).

Computerized adaptive testing (CAT) can be very efficient in comparison to traditional testing (Sands, Waters & McBride, 1997; Van der Linden and Glas, 2000; Wainer, Dorans, Flaugher,Green, Mislevy, Steinberg & Thissen, 1990; Weiss & Kingsbury, 1984). With a relatively short test length we already obtain a highly accurate ability estimate. This removes the objection to the use of a prior distribution. With an accurate test the weight of the prior in the final ability estimate is very small.

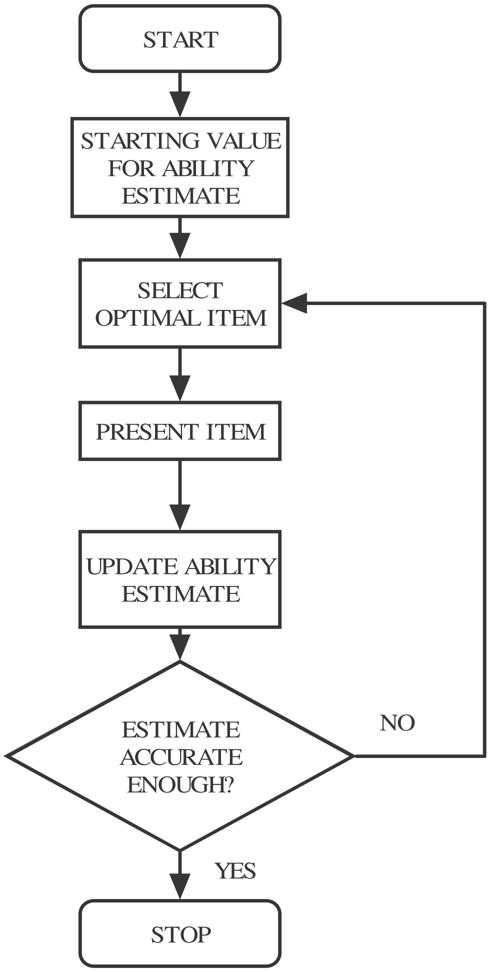


**Figure 9.3.** Flowchart of CAT with a stopping rule based on estimation acuuracy

In practice concessions have to be made in order to make CAT feasible. If we would use only items with maximum information given the estimated ability, then we would probably use a limited number of items from a large item pool frequently and other items would never be used. Several methods have been proposed to deal with this problem. (Revuelta & Ponsoda, 1998). See also Van der Linden & Reese (1998) for suggestions to some of the problems encountered in CAT.

When CAT is to be introduced a few aspects of testing with CAT must be attended to. Answering the CAT-test is different from answering a traditional test. The items must be answered consecutively; skipping items is not allowed. Therefore it is sound practice to study the validity of the test procedure. We should be alert to the possibility that the validity of the test changes with a change in procedure.

Recently, the interest in CAT is growing tremendously, especially because of its prospects in educational assessment. Also the future of psychological testing will be determined, among others, by CAT (see e.g. *Standards*, APA, AERA, & NCME, 1999; Bunderson, Inouye, & Olsen, 1989; Gregory, 2000, pp. 566-7; 569). In 1995 the American Council on Education published the *Guidelines for computer-adaptive test development and use in education*. More technical issues of CAT can be found in the special; spring issue 1997 of the *Journal of Educational Measurement*.

## 9.7 The Measurement of Change

The measurement of change is beset with problems:

- The first problem has to do with measurement error which gives rise to the phenomenon of the regression to the mean
- The second problem has to do with the question whether change can be interpreted as change along one dimension; can scores before and after a change be interpreted in terms of the same underlying construct?
- The third problem has to do with the limitations of the measurement instruments used. When the standing of a person on a latent scale increases, a higher score can be expected on the measurement instrument. A person with an intermediate score can have a large observed score gain. A person with a relatively high initial score cannot have a large observed gain: there is a maximum score on the test. In the measurement of change a ceiling effect is to be expected.

It is clear that IRT cannot solve all problems associated with the measurement of change. Using estimated scores on a latent scale defined by an IRT model can at least eliminate the problem of ceiling effects. An early proposal with respect to the use of IRT for the measurement of change has been made by Fischer (1976). Fischer did more than to propose to use the latent ability scale for measuring change, he also proposed to model the amount of change for each individual as a weighted sum of effects. Another Rasch model for the measurement of change over several occasions has been proposed by Embretson (1991).

## 9.8 Concluding Remarks

A recent canonization of IRT is the *Handbook of modern item response theory* (1997), edited by Wim J. van der Linden and Ronald K. Hambleton. Not only are many IRT models presented by experts in the field, but also examples are provided from educational and psychological (although limited in number) testing. Most of these models have been included in the previous two chapters of the present monograph. Interestingly, the models can be classified according to roughly the following criteria:

- response format (dichotomous, polytomous items; ordered versus unordered categorical data, open-end items);
- response time or number of successful attempts as responses on test items;
- unidimensional or multidimensional items in a test;
- type of response function (monotonous versus non-monotonous; type of the response function e.g. normal ogive, logistic, hyperbolic cosine, Cauchy);
- single versus multiple-group analysis.

Combining these five criteria maps a whole gamut of IRT models, giving work for a whole army of research workers for decades to come. And if the application of IRT models in psychological and educational testing is taken seriously, then not only is cooperation between basic and applied researchers a prerequisite, but also a selection of the applied fields necessary (e.g. performance assessment, test fairness, setting standards, certification, and the like in educational testing; the measurement of human abilities, measurement in personality, clinical and health psychology, developmental psychology, attitude measurement, personnel psychology). In the context of applied measurement also more attention should be paid to the interpretation of model parameters, in addition to more technical matters as model identification and parameter estimation.

But where are we now: what are the achievements and blessings of IRT? Is it a fair assessment by e.g. Goldstein & Wood (1989) or Blinkhorn (1997) that, to rephrase Horace, a mountain of IRT models gave birth to a silly little mouse of insight?

IRT has led to some fruitful results in the field of equating or research on the comparability of measures, on fairness in testing and test use (DIF research), and last but not least computerized adaptive testing (CAT). Specifically, in the field of CAT, IRT is indispensable. One of the major developments of educational and psychological testing in the fore-seeable future is CAT, and CAT is nigh to impossible without taking refuge to IRT. This development is predicted by Gregory (2000), among others, although he himself pays no attention to IRT in his much used textbook for advanced undergraduates in psychological testing. Together with another prediction that nationwide testing will be on the wane, this calls for diversified IRT item banking. Daniel (1999) argues that IRT is indispensable for improving the adaptive administration of intelligence tests.

It is not all roses, however, with IRT. One class of problems with IRT is methodological and technical in nature. For example, what does it really mean when assumptions are violated, and then, how to proceed? Surely, when the unidimensionality assumption is violated we can take refuge to multidimensional IRT. But what is the nature of those multiple dimensions, how to interpret them? How stable are the results on the item information function under deviations of ICC's from the normal ogive and logistic models (see Bickel, Buyske, Chang & Ying, 2001)? What does IRT contribute to test validation? Problems of a technical nature have to do with estimation of model parameters, and with model testing. The three-parameter logistic model, for example, requires sample sizes of at least 1000 for a moderate number of items (say 40) in a test to achieve stable parameter estimates. And with more elaborated models the estimation problems become more complicated.

A second class of problems has to do with the implementation of IRT for e.g. personality assessment. In personality assessment a wide variety of constructs exist. Can all these constructs be modeled appropriately by an IRT measurement framework? Reise (1999, p. 237-8) argues that most available IRT models are too restrictive for personality assessment.

Of course, it is no excuse of applied research workers and practitioners in educational and psychological testing to shun the use of IRT procedures because these are tedious to apply and difficult to understand. On the other hand, research workers would be ill advised if all of them should climb the IRT bandwagon, and leave classical and neoclassical (i.e. generalizability) test theory behind. These major test theories and their corresponding procedures for test development, validation and evaluation must continue to exist side by side, but cross-fertilization should be enhanced.

## Exercises

9.1 Two test items have been administered to a reference group R and a focal group F. The proportions correct are

$p_{1(\mathrm{R})} = 0.70$

$p_{1(\mathrm{F})} = 0.65$

$p_{2(\mathrm{R})} = 0.70$

$p_{2(\mathrm{F})} = 0.60$.

Is the second item biased against the focal group?

9.2 We have five 5 Rasch items with $b_1$ = -0.5, $b_2$ = -0.3, $b_3$ = 0.0, $b_4$ = 0.25 and $b_5$ = 0.5. We want to construct a two-item test that discriminates relatively well at $\theta_1$ = -0.5 and at $\theta_2$ = 0.5. Which combination of two items from the item bank with 5 items is best, given the criterion that the largest of the error variances at $\theta_1$ and $\theta_2$ should be as small as possible?

9.3 In an item bank we have items conforming to the 2PL model. The items have the item parameters: $b_1$ = -0.5, $a_1$ = 1.0, $b_2$= -0.25, $a_2$ = 2.0, $b_3$ = 0.0, $a_3$ = 0.7, $b_4$ = 0.25, $a_4$ = 1.0, $b_5$ = 0.5 and $a_5$ = 1.5. We test a person and the present point estimate of ability $\theta$ is 0.20. Which item should be presented next to this person?

# 10 TEST EQUATING

## 10.1 Introduction

In many situations in psychological and educational testing multiple forms of a test are made available to assess ability, achievement, performance or whatever. When persons are administered several test forms meant to measure the same ability, we want to be able to compare these persons' test scores. With parallel tests this can be done straightforwardly. Parallel tests measure the same content and share statistical specifications (equal means, standard deviations, and reliabilities). That is to say, scores on parallel tests are completely exchangeable. No comparison problem occurs with parallel forms of a test. More often than not multiple forms of a test that measure the same attribute are not parallel, and a comparison of scores is not straightforward, because test forms may differ in several respects (unequal means, unequal variances, unequal reliability, and the like). So, before comparing persons' or examinees' scores on multiple forms of the same test, it is necessary to establish, as nearly as possible, an effective equivalence between raw scores on the multiple forms of a test. This is the problem of equating.

Before going into the equating problem, we must keep in mind that in general the process of associating numbers with the performance of persons or examinees on tests is called scaling. This process leads to scale scores. In psychological and educational testing a number of scales is around (raw scores, normalized scores, stanines, and the like). The process of scaling must be distinguished from the process of equating. Equating procedures are used to insure that scores from the administration of multiple forms of a test can be used interchangeably. Test theorists and practicians differ in opinion, however, what conditions should be met for equated scores, i.e. the scores obtained after applying equating methods. Not only can interchangeability refer to alternative and weaker forms of strict parallelism of measurement instruments as discussed in Chapters 3 and 4, but also to test content and to the target population for which the test is intended. To be more precise, the following four conditions or properties of equated test scores are pertinent:

1. Same ability, i.e. alternative test forms must measure the same characteristic (ability, achievement, or performance).
2. Equity: for every group of persons or examinees given the same ability, the conditional frequency distribution of scores on one of the test forms, say test *Y*, is the same after transformation as the conditional frequency distribution of untransformed scores on the other test, test *X*.
3. Population invariance, i.e. the transformation is the same irrespective of the sample or group of persons from which it is derived.
4. Symmetry, i.e. the transformation is reversible, transforming the scores of form *X* to form *Y* is the same as transforming the scores of form *Y* to form *X*.

The explicit definition of condition 2 is given by Lord (1980, Chapter 13). If complete equity after equating or transformation of scores on test forms *X* and *Y* is observed, than both forms of the test are strictly parallel in the sense of classical test theory. Complete equity according to the definition by Lord is hardly feasible in practice, be it for the simple reason that very often reliabilities differ. Low ability examinees have an advantage with a relatively low reliability, whereas high ability examinees have an advantage with an accurate measurement of their ability, in other words, with a relatively reliable test. So, it should be clear how important it is to make tests as comparable as possible with respect to reliability.

After equating at least the expected score or true score on one test should be equal to the true score on the other test. In terms of true scores of two tests *X* and *Y* we should obtain

$$\mathrm{T}'_Y = \mathrm{T}_X,$$

in other words, after equating test *Y* with test *X* the true score on test *Y* is equal to the true score on test *X*. As already mentioned, meeting all four conditions or desirable equating properties is nigh to impossible. In actual equating practice they can only – hopefully as close as possible – be approximated. And as investigators in testing programs differ in opinion on what closeness of approximation entails, one or more of the conditions or properties are aimed at. Specifically, the conditions or desirable properties of equated scores are discussed in more detail by Lord (1980), Kolen (1999), and Petersen, Kolen & Hoover (1989).

Test equating is an empirical enterprise. It boils down to establishing a relationship between raw scores or scale scores in general on two or more test forms: data on multiple test forms have to be collected, and then appropriate equating methods have to be applied for transforming the scores. In Section 10.2 three basic equating designs for collecting data are outlined. In Sections 10.3 equipercentile equating is introduced, in Section 10.4 linear equating. Linear equating methods that make use of an anchor test are presented in Section 10.5. In Section 10.6 IRT-based equating methods are presented on an elementary level, without losing the gist and flavor of them, however. In the final Section 10.7 some concluding comments are made.

We mention some of the important publications here. The monograph by Kolen and Brennan (1995) appears to be the only recent book in the field of test equating. Angoff (1971; 1984 published as a separate monograph) gave one of the first extensive treatments, a chapter in *Educational Measurement*, 2nd Ed. In the third edition of *Educational Measurement* Petersen, Kolen and Hoover (1989) coined their contribution “Scaling, norming, and equating”. There is a special issue of *Applied Psychological Measurement* (Brennan, 1987), and a special issue of *Applied Measurement in Education* (Dorans, 1990) to give overviews of research and development in this field. More specifically, Skaggs & Lissitz (1986) review IRT equating, and Cook and Petersen (1987) discuss problems related to the use of conventional and IRT equating in less than optimal circumstances. Last, but not least, the new *Standards for educational and psychological testing* (APA, AERA, & NCME, 1999) should be mentioned. The importance and relevance of equating of tests is exemplified by including a special chapter, Chapter 4, on scales, norms, and score comparability or equating.

## 10.2 Some Basic Data Collection Designs for Equating Studies

There are several methods that can be used to equate scores on multiple test forms. These equating methods are tuned to the particular data collection design. Here we shall only discuss three basic data collection designs. A more extended list, including section pre-equating (Holland & Wightman, 1982) can be found in Petersen et al. (1989).

1) Design 1: Single-Group Design (Diagram 10.1.a)

In this design forms *X* and *Y* are given both to one group of persons or examinees. A disadvantage of this design is that much time is needed for the administration of the tests. Fatigue might play a role when persons answer the items of the second test.

Therefore, the best thing to do is to administer the tests in a different order to part of the persons. Technically speaking, this design is a counterbalanced random-groups design. The single group is split into two random-half samples and both half-samples are administered test forms *X* and *Y* in counterbalanced order, e.g. the first half-sample takes form *X* first, while the second half-sample takes form *Y* first. This is realized by administering the tests in rotation to a group of persons who are present at the test session.

2) Design 2: Random-Groups Design (Diagram 10.1.b)

In this design test forms *X* and *Y* are administered to different random samples from the population.

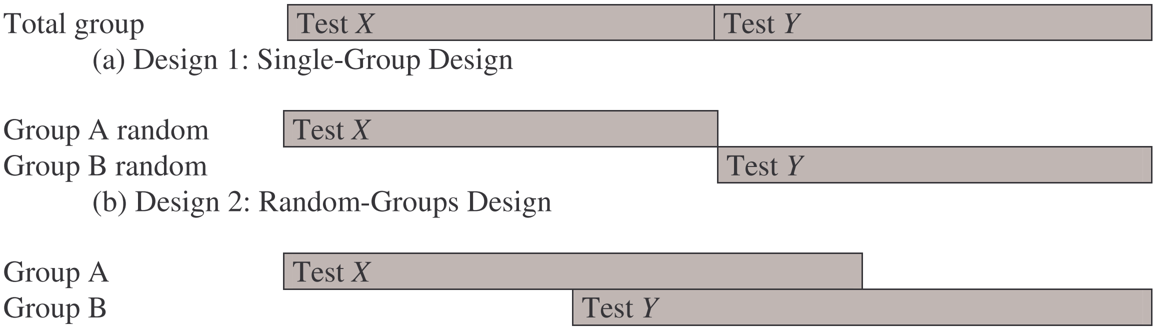
With large-scale examinations one of the tests, say test *X*, is the old test and the other test, test *Y*, is a new test to be equated to the old test. An equating study using Design 1 or Design 2 is not possible in this situation because the contents of the new test would become available prematurely. Design 3 does not have this disadvantage.

3) Design 3: Anchor-Test Design (Diagram 10.1.c)

In this design all persons are given a test *V* that is functionally equal to tests *X* and *Y*. So, test *X* and test *V* are administered to one sample of persons, and test *Y* and test *V* to another sample of persons. The two samples may differ from each other in a nonrandom way. The common test *V* is called the anchor test. Test *V* may be a common subtest of test forms *X* and *Y*; in that case we talk of an internal anchor test. Test *V* might also be a third test in which case we have an external anchor test. Tests *X* and *Y* can be related to each other by means of the common or anchor test *V*.

Design 3 can also be used with two random samples of persons. An advantage of Design 3 in comparison to Design 2 is that eventual differences between the two random groups of respondents can be corrected for. Lord (1955) proposed a statistical correction. For this correction it is not necessary that test *V* measures the same construct as tests *X* and *Y*, but, of course, the correction is better with a higher correlation between *V* and test forms *X* and *Y*.

The items from an internal anchor test should, of course, be insensitive to context. Also, the common items should have the same functionality in both test forms *X* and *Y*, i.e., the items need to behave similarly. Assume that test *X* has been administered a long time before test *Y*. Items from test *X* might have become obsolete and therefore have become more difficult at the time test *Y* is administered. Such obsolete items are not suitable as anchor items: inclusion in the anchor test would *not* result in equivalent scales after application of an equating procedure.



(a) Design 1: Single-Group Design

(b) Design 2: Random-Groups Design

(c) Design 3: Anchor-Test Design (internal anchor test)

**DIAGRAM 10.1.** Three basic data collection designs for equating studies

## 10.3 The Equipercentile Method

The property of population invariance of equating can only be approximated in practice, especially when raw scores on two test forms are used. So, it is important to define the population for which the relationship between two tests *X* and *Y* has to be obtained. In equipercentile equating raw scores on test forms *X* and *Y* are considered to be equated if they correspond to the same percentile rank in the population.

Suppose that we have administered two forms *X* and *Y* of a test to a large group of persons from the relevant population (Design 1). The tests have the same reliability and there are no context effects. Then two scores *x* and *y* are equivalent – apart from measurement error – if the two scores have an identical percentile rank, in other words if equal percentages of persons have these scores or lower scores on the tests. With Design 2 in principle the same definition of score equivalence can be used. The difference with Design 1 is that sample fluctuations introduce more error in the estimated relationship between the tests. The equating process for equating with the equipercentile method is demonstrated with Table 10.1 and Figures 10.1 and 10.2.

In Table 10.1 the percentile scores of two 40-item test forms *X* and *Y* are given. The percentile score associated with a particular raw score *y* on test form *Y* equals the percentage of persons with score $y - 1$ or a lower score plus half the percentage of persons with score *y*. In the table we find that raw score 20 on form *Y* corresponds to a percentile score equal to 30.3. In test form *X* raw score 20 corresponds to a percentile score equal to 45.3. The score on test form *X* that is equivalent to score 20 on test form *Y* must have the same percentile score: 30.3. There is no raw score on *X* with this percentile score. The equated score on *X* must have a value between 17 and 18. Linear interpolation gives the value 17.5 as the equated score on test *X*. In this way we can find equated scores *x* for all scores *y*. Those values are given in the table. Not in the table are the scores *y* equated to the raw scores 0 up to and including 40 on test form *X*.

The procedure of equipercentile equating of the raw scores on test forms *X* and *Y* is depicted in Figure 10.1. In Figure 10.2 the obtained relation between scores on test form *X* and scores on form *Y* is displayed.

**TABLE 10.1.** Percentile scores of two test forms *X* and *Y*, and scores *X* equated to scores *Y*

| raw | Percentile | percentile | score *X* | raw | percentile | percentile | score *X* |
|---|---|---|---|---|---|---|---|
| score | score *Y* | score *X* | equated to *Y* | score | score *Y* | score *X* | equated to *Y* |
| 0 | 0.0 | 0.0 | 4.0 | 21 | 35.3 | 51.6 | 18.3 |
| 1 | 0.0 | 0.0 | 4.0 | 22 | 40.6 | 57.5 | 19.3 |
| 2 | 0.0 | 0.0 | 4.0 | 23 | 45.3 | 63.9 | 20.0 |
| 3 | 0.0 | 0.0 | 4.0 | 24 | 50.7 | 69.8 | 20.9 |
| 4 | 0.0 | 0.0 | 4.0 | 25 | 56.9 | 74.2 | 21.9 |
| 5 | 0.0 | 0.1 | 4.0 | 26 | 62.4 | 78.0 | 22.8 |
| 6 | 0.0 | 0.1 | 4.0 | 27 | 67.2 | 81.4 | 23.6 |
| 7 | 0.2 | 0.5 | 6.3 | 28 | 72.2 | 85.0 | 24.5 |
| 8 | 0.6 | 1.2 | 7.1 | 29 | 77.1 | 88.3 | 25.8 |
| 9 | 1.0 | 2.0 | 7.7 | 30 | 81.7 | 90.8 | 27.1 |
| 10 | 1.5 | 3.3 | 8.3 | 31 | 85.7 | 93.0 | 28.2 |
| 11 | 2.2 | 5.0 | 9.2 | 32 | 89.5 | 95.6 | 29.5 |
| 12 | 3.3 | 6.8 | 10.0 | 33 | 92.8 | 97.3 | 30.9 |
| 13 | 4.7 | 9.2 | 10.8 | 34 | 95.2 | 98.3 | 31.9 |
| 14 | 6.5 | 12.7 | 11.8 | 35 | 97.5 | 99.1 | 33.2 |
| 15 | 9.1 | 16.9 | 13.0 | 36 | 99.1 | 99.5 | 35.1 |
| 16 | 12.2 | 21.6 | 13.9 | 37 | 99.6 | 99.8 | 36.2 |
| 17 | 15.9 | 27.5 | 14.8 | 38 | 99.7 | 100.0 | 36.7 |
| 18 | 20.7 | 33.4 | 15.8 | 39 | 99.9 | 100.0 | 37.3 |
| 19 | 25.8 | 39.0 | 16.7 | 40 | 100.0 | 100.0 | 38.0 |
| 20 | 30.3 | 45.3 | 17.5 | | | | |

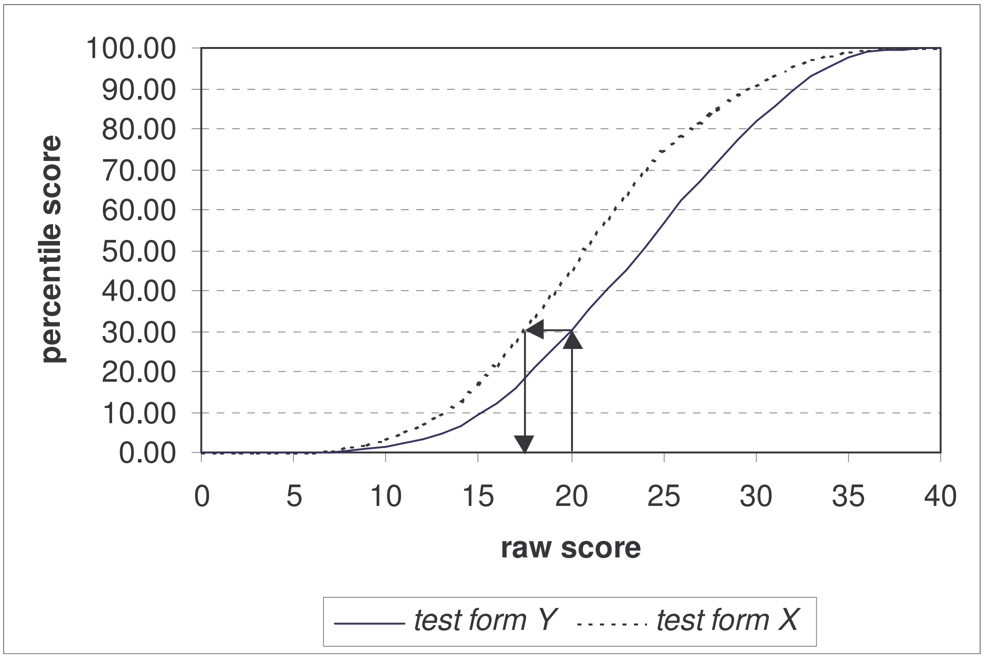


**FIGURE 10.1.** The percentile scores of test forms *X* and *Y*, and the construction of equated scores

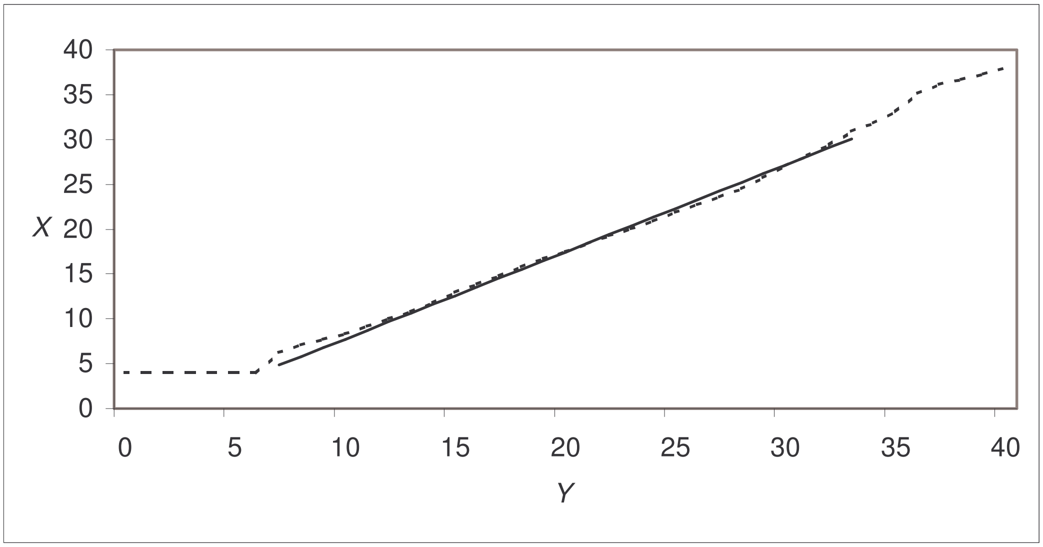


**FIGURE 10.2.** The equated scores on forms *X* and *Y*, obtained with the equipercentile method; also a linear approximation of the relation between the scores on *X* and *Y* is shown

Equipercentile equating is extremely sensitive to sampling fluctuations. This is especially the case at the low and high ends of the score scales where the computation of percentile scores rests on small numbers of observations. Several approaches have been suggested to diminish the influence of sampling fluctuations. All methods are based on smoothing. In presmoothing methods the score distributions of *x* and *y* are smoothed before equating the tests, in postsmoothing the obtained relationship between *X* and *Y* is smoothed (Kolen & Brennan, 1995, Chapter 3).

When test forms *X* and *Y* are approximately linearly related, the obvious smoothing technique is to equate tests by means of a linear equating method. As Figure 10.2 shows, linear equating might be adequate for a large range of scores.

The equipercentile method for equating can be applied to data obtained with all the designs discussed earlier in Section 10.2. How equating proceeds in practice is given by Kolen and Brennan (1995), among others. Also standard errors of equipercentile equating are presented there.

## 10.4 Linear Equating

If tests have about the same score distribution apart of their means and standard deviations, a linear equating method is sufficient. In linear equating a transformation is chosen such that scores on two test forms are considered equated if they correspond to the same number of standard deviations above or below the mean in the same group of persons. Because of the typical character of linear equating, we must take some measures for equating at high and low scores because linear equating inevitably leads to impossible transformed scores, negative scores and scores higher than the maximum score. For linear equating in Design 1 or Design 2 the following equation is used:

$$\frac{x-\overline{x}}{s_X}=\frac{y-\overline{y}}{s_Y}, \tag{10.1}$$

where *x* and *y* refer to the scores on the test forms to be equated. Using (10.1) we can write the score on *Y* after transformation to the scale of *X*, *y*′, as

$$y' = Ay + B, \tag{10.2}$$

where

$$A = \frac{s_X}{s_Y} \tag{10.3}$$

and

$$B = \bar{x} - A\bar{y}\,. \tag{10.4}$$

In Design 2 two random samples of persons are needed. Test form *X* is administered to one of these samples, group A. The mean and standard deviation of scores on *X* are obtained from this group. Test *Y* is administered to the second sample, group B. Group B gives a mean and standard deviation for scores on *Y*. If groups A and B cannot be regarded as random samples from the same population, equating is not possible with Design 2. An anchor test is needed to make a correction for group differences possible, in which case the data collection design for linear equating of scores is Design 3.

## 10.5 Linear Equating with an Anchor Test

In Design 3 we have an internal or external anchor test *V* that has the same function as test forms *X* and *Y*. By means of the common anchor test tests *X* and *Y* can be equated. First, we must define the population for which the equating relation is to hold. This population, the so-called *synthetic population*, might be defined by combined group A + B, but also other population definitions are possible (Kolen & Brennan, 1995, p. 106, 111). Using data for test *X* and *V*, and test *Y* and *V*, we can estimate means and standard deviations of *X* and *Y* in the synthetic population. Next, (10.2) through (10.4) are used to define the equivalence relationship between tests *X* and *Y*. Several methods for estimating means and standard deviations of tests *X* and *Y* in the synthetic population are available. Tucker proposed a method for linear equating that can be used if groups A and B do no differ very much in ability. The resulting equating equation is formally identical to a result obtained by Lord (1955) under the assumption of random groups A and B. Levine developed two methods for samples that may differ widely in ability. The first method can be used with equally reliable tests. The second method is suitable when the test forms differ in reliability. In the second case the true scores can be scaled to the same scale, but obviously raw scores cannot be equated.

The first method is described in Exhibit 10.1. The second method is easier to derive than the first Levine method or the Tucker method. Due to the fact that the second Levine method is defined for true scores, the computation of means and standard deviations of test scores for the synthetic population can be avoided.

### Exhibit 10.1. Levine's first method: equally reliable tests

With equally reliable tests the mean and standard deviation of test forms *X* and *Y* are estimated for a synthetic group. Here, we estimate the mean and variance of *X* and *Y* for the total group T = A + B. The procedure is illustrated for test form *X*.

Two assumptions are made with respect to test form *X* and the common test *V*. The first assumption is that the true scores of *X* and *V* are linearly related. This assumption leads to two equations:

a) $\overline{\mathrm{T}}_{X(\mathrm{T})} - \dfrac{s_{\mathrm{T}_X(\mathrm{T})}}{s_{\mathrm{T}_V(\mathrm{T})}} \overline{\mathrm{T}}_{V(\mathrm{T})} = \overline{\mathrm{T}}_{X(\mathrm{A})} - \dfrac{s_{\mathrm{T}_X(\mathrm{A})}}{s_{\mathrm{T}_V(\mathrm{A})}} \overline{\mathrm{T}}_{V(\mathrm{A})}$

and

b) $\dfrac{s_{\mathrm{T}_X(\mathrm{T})}}{s_{\mathrm{T}_V(\mathrm{T})}} = \dfrac{s_{\mathrm{T}_X(\mathrm{A})}}{s_{\mathrm{T}_V(\mathrm{A})}}$,

i.e. the intercept and the slope of the relation of the true scores of *X* and *V* is the same in groups A and T.

The second assumption is that the variance of measurement errors for test form *X* is the same in group A and group T,

c) $s^2_{X(\mathrm{T})}(1 - r_{XX'(\mathrm{T})}) = s^2_{X(\mathrm{A})}(1 - r_{XX'(\mathrm{A})})$.

Using a and b and substituting the observed mean for the true score mean we obtain

$$\bar{x}_{\mathrm{T}} = \bar{x}_{\mathrm{A}} + \frac{s_{\mathrm{T}_X(\mathrm{A})}}{s_{\mathrm{T}_V(\mathrm{A})}}(\bar{v}_{\mathrm{T}} - \bar{v}_{\mathrm{A}}).$$

Using b and c we obtain

$$s^2_{X(\mathrm{T})} = s^2_{X(\mathrm{A})} + \frac{s^2_{\mathrm{T}_X(\mathrm{T})}}{s^2_{\mathrm{T}_V(\mathrm{T})}}(s^2_{V(\mathrm{T})} - s^2_{V(\mathrm{A})}).$$

The observed-score variance of test form *X* in the total group can be obtained when the ratio between the true-score variance of *X* and the true-score variance of *V* is known. An estimate of this ratio is presented in the main text.

Next, the mean and variance of test form *Y* in the total group are estimated. Finally, Equations 10.2 through 10.4 are used to equate test forms *X* and *Y*.

In the second Levine method it is assumed that the true scores on *X* and *V* are linearly related, and similarly that the true scores on *Y* and *V* are linearly related. A true score on test *X*, $\mathrm{T}_X$ is equivalent to a true score on test *V*, $\mathrm{T}_V$ if the two scores have the same *z*-score within the same group, say group A, of persons:

$$\frac{\mathrm{T}_X - \mu_{\mathrm{T}_X(\mathrm{A})}}{\sigma_{\mathrm{T}_X(\mathrm{A})}} = \frac{\mathrm{T}_v - \mu_{\mathrm{T}_V(\mathrm{A})}}{\sigma_{\mathrm{T}_V(\mathrm{A})}}. \tag{10.5}$$

This equation can be rewritten as:

$$\mathrm{T}_X = \gamma_{XV}(\mathrm{T}_V - \bar{v}_{(\mathrm{A})}) + \bar{x}_{(\mathrm{A})}, \tag{10.6}$$

where observed-score means are substituted for true-score means. In the equation $\gamma_{XV}$ denotes the ratio between the true-score standard deviation on *X* and the true-score standard deviation on *V*. This ratio is assumed to be group-independent. A similar equation can be obtained for the relation between true scores on *Y* and true scores on *V*. Coefficients for the equation relating *Y* and *V* can be obtained from group B:

$$\mathrm{T}_V = (\mathrm{T}_Y - \bar{y}_{(\mathrm{B})}) / \gamma_{YV} + \bar{v}_{(\mathrm{B})}. \tag{10.7}$$

Substitution of $\mathrm{T}_v$ from (10.7) in (10.6) produces the following relationship between the true-score scales of tests *X* and *Y*:

$$\mathrm{T}_X = \frac{\gamma_{XV}}{\gamma_{YV}}[\mathrm{T}_Y - \bar{y}_{(\mathrm{B})}] + \gamma_{XV}[\bar{v}_{(\mathrm{B})} - \bar{v}_{(\mathrm{A})}] + \bar{x}_{(\mathrm{A})}. \tag{10.8}$$

Next raw scores on *X* and *Y* are equated as if they were true scores. The correction for the difference in ability level between groups A and B is one of the differences with (10.2) – (10.4). The second difference has to do with $\gamma$. Angoff (1971) called the ratio of true-score standard deviations $\gamma$ effective test length. He assumed that test *X* can be regarded as a combination of $\gamma_{XV}$ tests parallel to anchor test *V*. Similarly, test *Y* might be regarded as a test composed of $\gamma_{YV}$ tests parallel to test *V*. This is a stronger requirement than the requirement that the three tests *X*, *Y* and *V* have linearly related true scores. With Angoff's assumption the coefficients $\gamma$ can easily be determined. In case test *V* is included in test *X* $\gamma_{\mathrm{XV}}$ is computed as:

$$\gamma_{XV} = \frac{s_x}{\mathrm{s}_v r_{xv}}. \tag{10.9}$$

So, in this case factor $\gamma$ equals the inverse of the regression coefficient for the regression of *V* on *X* (Angoff, 1953). The coefficient is estimated from responses in group A. The factor $\gamma_{YV}$ is estimated from the responses in group B. If *V* is an external test, another equation than (10.9) is needed. Standard errors for the Levine procedure are given by Hanson, Zeng & Kolen (1993).

When the common test *V* has not the same function as test forms *X* and *Y*, equating is possible when the two groups A and B are random samples, using a method proposed by Lord (1955). If this is not the case, equating is not possible, but it is still possible to obtain comparable scores for test forms *X* and *Y*. Scores on *X* and *Y* might be defined as *comparable* if they are predicted by the same score on *V*. The definition of comparable scores as scores that are predicted by the same score on a third test, is not the only definition possible. There are other definitions of comparability. The issue of comparability of scores is discussed at some length by Angoff (1971, p. 590-597).

## 10.6 IRT Models for Equating

Instead of the equipercentile method or the linear method of equating a method that is based on IRT can be used. Equating with IRT has a large advantage over equating with the classical approach. With an IRT model that fits the nonlinearities inherent in equating do not present a problem. IRT models can be used in *horizontal equating* as well as in *vertical equating*. In horizontal equating different tests are meant for persons of similar abilities; equating as discussed so far is horizontal equating. In vertical equating tests are constructed for target groups of different ability levels. The difference in test difficulty is planned but for score interpretation scores should be brought to the same scale. It is still necessary that all items are relevant for all examinees. Equating is not achieved if younger examinees have not been exposed to material tested in the unique items of the higher level test tailored to the ability of a group of older examinees (Petersen et al., 1989). It should also be clear that in vertical equating tests are not equated in the sense that they may be used interchangeably after equating.

In principle three equating approaches for two test forms $X$ and $Y$ sharing a common set of items are possible within the IRT context (Petersen, Cook & Stocking, 1983):

A Simultaneous scaling. The item parameters of both tests are estimated jointly in one analysis. For this approach we need software that allows for incomplete data: each person has answered only a subset of all items.

B The responses to tests $X$ and $Y$ are analyzed separately. In the analysis of the second test the item parameters of the common items are fixed to their values obtained in the analysis of the first test. The scales of $X$ and $Y$ can be related to each other by means of the scale values of the common items.

C The responses to tests $X$ and $Y$ are analyzed separately. The difference with approach B is that the parameter values of the common items are not fixed to their values obtained in the analysis of the first test. Again the scales of test $X$ and $Y$ can be related to each other by means of the scale values of the common items.

When approach A is chosen and MML is the estimation method, characteristics of the latent ability distributions involved should be allowed to differ. Alternative C seems easiest to implement. Let us consider this approach in the context of the three most popular IRT models: the Rasch model, 2PL model and 3PL model.

*The Rasch Model*

With the Rasch model the third approach is very straightforward. We need the averages of the $b$ parameters of the common items in test $X$ and in test $Y$. Suppose that we have $k$ common items with the following averages:

$$\overline{b}_{X(\mathrm{c})} = \frac{1}{k}\sum_{i=1}^{k} b_{i\,X(c)}, \overline{b}_{Y(\mathrm{c})} = \frac{1}{k}\sum_{i=1}^{k} b_{i\,Y(c)} \qquad (10.10)$$

The $b$ parameters of both tests would be on a common scale if the average parameter value for the common items would be equal for both tests. So, the $b$ values and θ values of test $Y$ can be brought on the same scale as those of test $X$ with the transformation:

$$b_i^* = b_i - \overline{b}_{Y(\mathrm{c})} + \overline{b}_{X(\mathrm{c})}, \theta^* = \theta - \overline{b}_{Y(\mathrm{c})} + \overline{b}_{X(\mathrm{c})} \qquad (10.11)$$

We do not have the parameter values of the difficulty parameters, but only estimated values. The estimated values are not equally accurate. So we might consider using weighted averages instead of the unweighted average in (10.11). Such a method has been proposed by Linn, Levine, Hastings, & Wardrop (1981).

### *The 2PL Model*

In the 2PL model equating is a bit more complicated because the parameters are defined on an interval scale. The common items can have different *a* values as well as different *b* values. Because of the interval character of the latent scale *b*-parameter values of the common items of test *Y* are linearly related to the values for test *X*:

$$b_{i(X)} = db_{i(Y)} + e\,, \tag{10.12}$$

and the values of the *a* parameters are related through:

$$a_{i(X)} = {}^{a_{i(Y)}}\!/_{d}\,. \tag{10.13}$$

The coefficients *d* and *e* must be obtained in order to bring the parameters of the common items, and consequently the parameters of all items, to the same scale.

The simplest solution is to find the transformation by which the average *b* value of the common items and the standard deviation of the *b* values of the common items is equal in both tests. This is the so-called ‘mean and sigma’ method. With this method the value of *d* is

$$d = \frac{s_{b_{X(\mathrm{c})}}}{s_{b_{Y(\mathrm{c})}}}, \tag{10.14}$$

i.e. the ratio of the standard deviation of the common *b*-values in test form *X* to the standard deviation of the common *b*-values in test form *Y*, and the value of *e* is:

$$e = \overline{b}_{X(\mathrm{c})} - d\overline{b}_{Y(\mathrm{c})}\,. \tag{10.15}$$

A robust alternative is the previously mentioned weighted method proposed by Linn et al. (1981).

We have two sets of parameter estimates for the common items. One set is computed along with the other item parameters in test *X*. The other set is computed along with the other item parameters in test *Y*. We also can compute two test characteristic curves - the sums of the ICCs of the items in the tests - for the subset of common items. After test equating these two test characteristic curves should be very similar. In the characteristic curve methods (Haebara, 1980; Stocking & Lord, 1983) coefficients *d* and *e* are obtained for which these test characteristic curves are as similar as possible.

### *The 3PL Model*

The 3PL model is also defined on an interval scale, but the presence of a pseudo-chance-level parameter *c* complicates the equating of tests. When we analyze two tests *X* and *Y* separately, the estimated *c* of a common item can have one value in test *Y* and another in test *X*. This difference is related to differences in the other parameter estimates of the particular item. The errors in the pseudo-chance-level parameters can have a disturbing effect on the relationship between the *b* parameters and the *a* parameters of the common items. In other words, we may expect a disturbing influence on the linear relationship between the item difficulties in test *X* and those in test *Y*. Equating tests *X* and *Y* is not simply achieved by using a linear transformation for the *b* values of the common items. With the 3PL model more steps are needed. In a preliminary analysis we obtain estimates of the parameters *c*. For a common item one value for *c* is chosen on basis of the two different values obtained in the analyses of tests *X* and *Y*. The chosen value can be the average of the two estimates. In the final analysis the *c* parameter of a common item is fixed to this value for both tests.

After scaling the two test forms on a common latent scale the relation between the true scores of both test forms can be computed. For each value of θ the corresponding true score of test form *X* and the corresponding true score of test form *Y* can be computed:

$$\tau_X = G(\theta) = \sum_{i \in X} P_i(\theta) \tag{10.16}$$

and

$$\tau_Y = H(\theta) = \sum_{i \in Y} P_i(\theta)\,. \tag{10.17}$$

The two true scores corresponding to the same θ are equated with the following formula:

$$\tau_X = G(H^{-1}(\tau_Y))\,, \tag{10.18}$$

i.e., we take the true score on test form *Y*, compute the corresponding value of θ, and, next, compute the true score on form *X* for this value of θ. True-score equating does not work in the 3PL model for equating observed scores below the chance level. One obvious procedure to obtain the relation between the tests below the level of the pseudo-chance level is to use (linear) interpolation. Lord (1980) suggested an alternative, a raw-score adaptation of the IRT-equating method. In this procedure the distribution of θ is estimated for some group. Given this distribution the marginal distributions of *x* and *y* can be estimated. Next, *X* and *Y* can be equated through equipercentile equating! Of course the outcome depends to some extent on the group.

*Other models*

The equating methodology can be extended to the linking of tests with polytomous items. Cohen and Kim (1998) present an overview of linking methods under the graded response model. This model sometimes is used in connection with the judgment by raters of constructed responses. The fact that judges play a role complicates the linking process. For, it is by no means sure that judges have, for example, a stable year-to-year severity of judgment (Ercika, Schwartz, Julian, Burket, Weber & Link, 1998; Tate, 1999).

## 10.7 Concluding Remarks

At a certain moment a new test form *Y*, equated with an old test form *X*, is replaced on its turn by a more recent test form *Z*. So, we are not ready after equating the two test forms. After some time we have obtained a chain of multiple test forms equated to each other. With more than two test forms we can use alternative equating designs. In Figure 10.3 the three possible designs with three different tests *X*, *Y* and *Z* are displayed. For more information on flexible equating designs, see Exhibit 10.2.

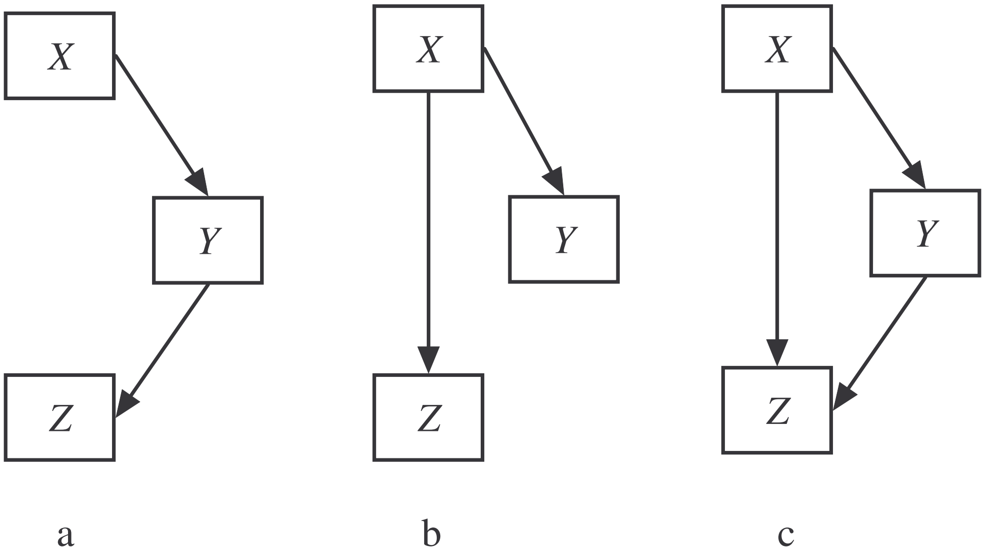


**FIGURE 10.3.** The three equating designs with three test forms *X*, *Y* and *Z*

**Exhibit 10.2. A History of Equating**

The Test of English as a Foreign Language (TOEFL) has had a history of a frequent introduction of new test forms. These tests should yield scores that can be used exchangeable. Equating has been used to eliminate possible differences between the test forms. First, conventional equating methods were used. From 1978 equating tests has been based on IRT, using the three-parameter logistic model.

Cowell (1982) describes the history of this change in practice. He also notes the change that IRT-equating makes possible. With IRT-equating data can be obtained from different previous tests. This allows us to try out items in pretests given to relatively small groups of persons. This means an improvement in test security.

With IRT-equating it is possible to have a new test that entirely consists of previously administered items with item characteristics estimated on a common scale, i.e. items selected from an item bank. This 'pre-equating' of items fastens the process of scoring.

If tests can be made from an existing pool of items, adaptive testing comes into reach, a fact also mentioned by Cowell. Nowadays two sections of the computer-based version of the TOEFL are adaptive.

One alternative is to equate test form *Y* with test form *X* using a common subtest of *Y* and *X*, and next test form *Z* with *Y* using a common subtest of *Z* and *Y* (Figure 10.3a). Test *Z* has no items in common with *X*, but test *Z* and test *X* are on the same scale as well through test *Y*. A weakness of the design is the risk that the inevitable equating errors cumulate. That risk is avoided in the design of Figure 10.3b. In this design both test forms *Y* and *Z* are directly equated with form *X*. Herewith test form *X* has become a standard. This solution has a disadvantage too. Items from *X* may become obsolete. Items from *X* might also become known; so many examinees can learn these items by heart. The third design (Figure 10.3c) is more balanced. In this design test form *Z* has common items with both forms *X* and *Y*. We may verify whether direct equating of test forms *Z* and *X* results in the same score transformation as equating forms *Z* and *X* through test *Y*. So, this design has a control mechanism. Another advantage of this design in comparison with the design in Figure 10.3b is that a limited number of items from test form *X* is needed in later test forms.

When IRT is used an adequate choice of the model is very important. If guessing occurs or is highly likely, then we prefer to use a model with a pseudo-chance-level parameter, otherwise equating errors are bound to be made, especially in the low score range and with vertical equating.

In this chapter we took the position that equating is necessary. However, whether one should equate test forms is something that has to be decided. In the final exhibit, Exhibit 10.3, this problem is discussed.

**Exhibit 10.3. To Equate or Not to Equate**

How to decide whether to equate multiple test forms or not? So far no systematic study answering this question has been performed. It is clear that equating of test scores is a prerequisite for large programs of educational and psychological testing where multiple test forms are involved of e.g. scholastic aptitude. The criteria to decide whether or not to equate forms depend at least on the following objectives:

- choosing the optimal design for data on the forms to be equated
- selecting the best equating procedure that meets the conditions or desirable properties for equating.
- minimizing statistical errors when conducting an equating study

These objectives, however, are intertwined. Minimizing statistical errors, systematic as well as random, depends on the study design and the equating procedure used. In design considerations sample size plays a crucial role and sample size affects statistical error. To be sure, minimizing equating error is a major goal when deciding whether or not to equate forms, when designing equating studies, and when conducting equating (Kolen, 1999, p. 171).

No conclusive answer on the question of this exhibit can be given. Kolen (1999, pp. 171-4) has a number of relevant comments to make on the equating methodology to date. He also briefly describes some testing programs with equating studies. Of course, much can be learnt from the relevant research reports on testing programs where equating is a must, or at least an essential standard.

## Exercises

10.1 We have two tests forms $X$ and $Y$. In a large random selection of persons the mean score on $Y$ is 60.0 and the standard deviation is 16.0. In a second random selection the mean score on $X$ is 55.0 and the standard deviation is 17.0. We want to equate $Y$ with $X$. With which score on $X$ does the score $y = 50$ correspond according to the linear equating method?

10.2 We have two groups of persons. One group is administered test form $Y$, the other is administered test form $X$. Suppose we know that the group that answered test form $Y$ is somewhat better than the other group. What can you say about the $x$ score equivalent to $y = 50$ in exercise 10.1? Explain your answer.

10.3 $V$ is a subtest of test $X$. Assume that test $X$ is the sum of subtest $V$ and $k$ - 1 tests parallel to subtest $V$. Prove that $k$ (= $\sigma_{T(X)}/\sigma_{T(V)}$) = $\sigma^2_X/\sigma_{XV}$ (Equation 10.9).

10.4 We have two test forms, $X$ and $Y$, each with five items. Both tests are analyzed with the Rasch model. The items are numbered consecutively in the table below. We notice that two items, items 3 and 5, are common to X and $Y$.

| item | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| *b* test *Y* | -1.5 | -1.0 | 0.0 | 0.5 | 2.0 | | | |
| *b* test *X* | | | -0.5 | | 1.5 | -0.5 | -0.5 | 0.0 |

Compute the estimated item parameters of the items of *Y* on the scale defined by test form *X*. Give the relationship between true scores for *X* and *Y* for $\theta$ = -4.0 (0.5) 4.0.

10.5 In a study test *X* has been administered to a group of high ability examinees, and test *Y* to a low ability group. Both groups also have been tested with test *V*. In a second study all examinees have been tested with test *V*. The examinees with relatively high scores on *V* have been tested with test *X*, the other examinees have been tested with test *Y*. All three tests are supposed to measure the same characteristic. Test *X* is relatively difficult, and test *Y* relatively easy, but we do not know how much the tests differ in difficulty level. What is the characteristic difference between the two studies and what are the consequences of this difference?

# ANSWERS

1.1 Researcher A probably will obtain a lower gain for the best pupils than researcher B. On the raw score scale the score cannot exceed the maximum score. In the study by researcher A there is a ceiling effect.

1.2 Imagine what might happen to the ranks if a player was added. The score scale is a typical example of an ordinal scale.

2.1 The test center compares the persons with other persons who have been tested at the same moment of the day. For the test center a fixed moment is part of the definition of the true score. The persons that are tested may have another point of view. They might compare their results with those of other persons who have been tested by other centers at different moments. In that case one should generalize over accidental variation in test administration times when defining true score. If the outcome of a test is notably influenced by the moment of the day on which the test is administered, it is relevant to know in which way moment of the day was dealt with in the norming study by the test developer.

2.2 It is possible that persons get used to the test format or that fatigue plays a role when the persons make the second test. In that case the condition of experimental independence is not satisfied. In a study with two tests A and B the role of fatigue can be controlled to a certain extent by using a test design in which one group of persons is administered test A first and next test B, and another group of persons is administered the tests in reverse order.

3.1 Application of Formula (3.3) gives 5.0.

3.2 The result of the computations is presented in the table below. The reliability increases strongly at first. For larger values of *k* the increase becomes smaller. For large values of *k* the reliability approaches the limiting value 1.

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_{xx'}$ | 0.50 | 0.67 | 0.75 | 0.80 | 0.83 | 0.86 | 0.88 | 0.89 | 0.90 | 0.91 | 0.92 | 0.92 | 0.93 | 0.93 | 0.94 | 0.94 |

3.3 The standard error of estimation equals the standard error of measurement times the square root of the reliability coefficient. For $\rho_{xx'} = 0.5$ the square root equals 0.71, for $\rho_{xx'} = 0.9$ it equals 0.95. So, for $\rho_{xx'} = 0.5$ the ratio of standard errors equals 0.71, for $\rho_{xx'} = 0.9$ the ratio equals 0.95. The Kelley estimate of true score is equal to 35.0 for $\rho_{xx'} = 0.5$, and equal to 31.0 for $\rho_{xx'} = 0.9$. For low reliability confidence intervals for the true score based on the observed score and the standard error of measurement deviate strongly from confidence intervals obtained using the Kelley point estimate of true score and the standard error of estimation. For high reliabilities the difference between the two approaches is relatively small.

3.4 Use the formula for the correction for attenuation. From this formula it can be deduced that

$$\rho_{XY} \leq \sqrt{\rho_{XX'}}\sqrt{\rho_{YY'}} \leq \sqrt{\rho_{XX'}}\ .$$

Clearly the correlation with a criterion cannot exceed the square root of the reliability. In other words, the correlation of the observed scores with their true scores gives an upper limit to the correlation of a measurement instrument with other variables. The maximum correlation for a reliability equal to 0.49 is 0.7.

3.5
$$\rho_{X(k)Y} = \frac{\text{cov}(X(k),Y)}{\sigma_Y \sigma_{X(k)}} = \frac{\text{cov}[(k\text{T}_X + \sum E_{X(i)}),Y]}{\sigma_Y \sqrt{k^2\sigma^2_{\text{T}_X} + k\sigma^2_{E_X}}}$$

$$= \frac{k\,\text{cov}(\text{T}_X,Y)}{\sigma_Y \sqrt{k^2\rho_{XX'}\sigma^2_X + k(1-\rho_{XX'})\sigma^2_X}} = \frac{k\rho_{XY}}{\sqrt{k^2\rho_{XX'} + k(1-\rho_{XX'})}}\,.$$

When $k$ goes to infinity, the formula can be simplified to

$$\rho_{X(\infty)Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}}}$$

(assuming $\rho_{XX'} > 0$).

4.1 The variance of total scores – the sum of entries in the table - equals 89.0. The sum of the variances of the subtests equals 37.0. Coefficient $\alpha$ is equal to 0.88.

4.2 In order to be able to apply Formula (4.7) we have to determine the values of the factor loadings $a_1$, $a_2$ and $a_3$. The factor loading $a_1$ equals the square root of $(\sigma_{12}\sigma_{13})/\sigma_{23}$ (see Formula 4.6). The factor loading $a_2$ equals the square root of $(\sigma_{21}\sigma_{23})/\sigma_{13}$ and $a_3$ equals the square root of $(\sigma_{31}\sigma_{32})/\sigma_{12}$. The computation of these factor loadings results in: $a_1 = 2.0$, $a_2 = 3.0$ and $a_3 = 4.0$. The reliability according to Formula (4.7) is $9.0^2/89.0 = .91$, a value which is a bit higher than the reliability estimated with coefficient $\alpha$.

4.3
$$\alpha = \frac{n^2\text{ave(cov)}}{\sigma^2_X} = \frac{n^2\text{ave(cov)}}{n\text{ave}(\sigma^2_i) + n(n-1)\text{ave(cov)}} = \frac{n^2 r^*}{n + n(n-1)r^*} = \frac{n\rho}{1+(n-1)\rho},$$

where *ave* stands for *average*; $r^* = \text{ave(cov)}/\text{ave}(\sigma^2_i)$. Because the items are parallel $r^*$ is equal to the common value of the inter-item correlation $\rho$. The result is identical to the Spearman-Brown formula.

4.4 The correlation between $X_i$ ($i = 1, 2$) and an arbitrary third test $Y$ is given by

$$\rho_{X_iY} = \frac{\text{cov}(X_i,Y)}{\sigma_{X_i}\sigma_Y} = \frac{a_i}{\sigma_{X_i}}\frac{\text{cov}(\text{T},Y)}{\sigma_Y} = \sqrt{\rho_{X_iX_i'}}\frac{\text{cov}(\text{T},Y)}{\sigma_Y},$$

where T is the true score on the common true score scale defined by $\mu_\text{T} = 0.0$ and $\sigma^2_\text{T} = 1.0$. So, congeneric measurements have identical patterns of correlations with other variables.

4.5 a. The observed score variance $\sigma^2_D$ is equal to $\sigma^2_X + \sigma^2_Y - 2\sigma_X\sigma_Y\rho_{XY}$. This gives us the value 9.6 for the observed variance of the difference scores. The true score variance $\sigma^2_{\text{T}(D)}$ is equal to $\sigma^2_{\text{T}(X)} + \sigma^2_{\text{T}(Y)} - 2\sigma_{\text{T}(X)}\sigma_{\text{T}(Y)}\rho_{\text{T}(X)\text{T}(Y)}$. The true score variances are obtained by multiplying the observed score variances with the reliabilities. The covariance of the true differences, $\sigma_{\text{T}(X)}\sigma_{\text{T}(Y)}\rho_{\text{T}(X)\text{T}(Y)}$, is equal to the covariance of the observed differences,

$\sigma_X\sigma_Y\rho_{XY}$. This gives us the value 3.2 for $\sigma^2_{T(D)}$. The reliability of the difference scores is low: 3.2/9.6 = 0.33. The low value is due to the high correlation between the true scores of the two tests.

b. The variance of the differences is 9.6. This is much larger than 6.4, the value of the error variance obtained from (4.19). The variance of the differences is larger because of the fact that the true scores on *X* and *Y* differ.

4.6 The condition of experimental independence might be violated. This can affect the reliability estimate as well as reliability itself. The violation of experimental independence can be eliminated through a redefinition of the items. Items belonging together might be treated as a single item when reliability is estimated from responses to the items. This redefinition of the item level might produce a new problem. The true score variance of an item consisting of many subitems can be much larger than the true score variances of other items. In this case $\alpha$ would underestimate the reliability. The effect of large differences between true score variances can be avoided by grouping all items in item clusters before reliability is estimated.

4.7 The inequality of means and correlations indicates that the tests are not parallel or tau-equivalent. The equal covariances indicate equal true-score variances. So, the three tests are essentially tau-equivalent.

4.8 When no specific effect like a learning effect is expected, we may assume that the true scores on both occasions are equal. The expected observed score on both occasions equals the true score. The expected true score on the second occasion given an observed score equal to 30 at the first occasion is 35.0, under the assumption that the regression of true score on observed score is linear (application of the Kelley formula). The expected difference score equals 5 (35 – 30). This is the so-called regression effect.

4.9 The true-score variance of the composite is 0.8 × 25.0 + 0.6 × 25.0. The observed-score variance of the composite is 25.0 + 25.0. The reliability is 0.7; see also Equation 4.11. For the reliability of the lengthened test we use the Spearman-Brown formula. We obtain 0.824 as the reliability of the lengthened test. We see that the reliability of a test composed of noncorrelating subtests can be high. Of course, it makes no sense to combine noncorrelating subtests into one test.

5.1 In the computation of the variance we divide the numerator by the number of persons minus one. The item variances are: 0.1342, 0.2211, 0.2395, 0.1974, 0.2211, 0.2605, 0.2632, 0.2395, 0.2605 and 0.2632. The sum of the variances equals 2.300. The variance of the total scores is 4.011. Coefficient $\alpha$ is (10/9)(1- 2.300/4.011) = 0.47.

The results of the analysis of variance are given in the following table:

| Source of variation | Sum of squares | Degrees of freedom | Mean squares |
|---|---|---|---|
| Persons | 7.620 | 19 | 0.4011 |
| Items | 2.920 | 9 | 0.3244 |
| Residual | 36.080 | 171 | 0.2110 |
| Total | 46.620 | | |

The variance component for persons equals (0.4011 - 0.2110)/10 = 0.019. The variance component for items equals 0.006. The residual (0.211) is by far the largest component, more than ten times as large as the variance component for persons. Measurement errors as well as the interaction between persons and items are part of this component. Due to the fact that there are no replications, the error and interaction components cannot be separated. The interaction component must be larger than zero because the items, which differ in difficulty level, cannot be essential tau-equivalent measurements of the underlying trait.

The generalizability coefficient for a test of ten items is equal to 0.47. This value is the same one as the value of coefficient $\alpha$ (as it should be: the coefficients are mathematically identical).

The example was chosen only to keep the computational burden low. It should be clear that the estimated variance components are unreliable due to the small number of persons and items.

5.2 The estimated residual component equals 0.65. The variance component for the interaction items × judges $\sigma^2_{ij}$ is (45.65 - 0.65)/500 = 0.090. The (estimated) variance component $\sigma^2_{pj}$ is equal to 0.010, $\sigma^2_{pi}$ equals 1.500, $\sigma^2_{j}$ equals 0.050, $\sigma^2_{i}$ equals 0.50 and $\sigma^2_{p}$ equals 0.175. The residual component is the largest component. The variance components involving judges are relatively small: judges seem to be reasonably well exchangeable. The generalizability coefficient for 15 items and 4 judges equals 0.61 (Formula 5.11 or 5.12).

5.3 See Formula (5.14).
a) $n'_i = 30$, $n'_j = 4$, the estimate of $E\rho^2$ is 0.75,
b) $n'_i = 60$, $n'_j = 2$, the estimate of $E\rho^2$ is 0.83.
In (a) as well as in (b) the total number of observations per person is twice the number used in the generalizability study. An increase of the number of items has a strong effect on generalizability, even if the number of judges decreases. One might have expected this result in view of the outcome of exercise 5.2. The variance components in which judges are involved, are relatively small; the judges are relatively well exchangeable.

5.4 The number of observations for a particular combination of $p$ and $i$ is $n_j$. This is the coefficient of the variance component for the interaction of $p$ and $i$:
$a = c = e = n_j$.
The other coefficients are
$b = n_in_j$ and $d = n_pn_j$.

5.5 When the correlation between judges is computed, the test items are regarded as fixed. The correlation can be written as the generalizability coefficient:

$$E\rho^2_{\text{Rel}} = \frac{[\sigma^2_p + \sigma^2_{pi}/n_i]}{[\sigma^2_p + \sigma^2_{pi}/n_i] + ([\sigma^2_{pj} + \sigma^2_{pij}/n_i] + \sigma^2_e/n_i)/n_j},$$

with $n_j$ equal to 1 (Maxwell & Pilliner, 1968).

5.6 The relative error variance is the error variance that plays a role in the generalizability coefficient for the crossed $p \times i \times j$ design. The error variance is equal to

$$\sigma^2_{\text{Rel}} = \sigma^2_{pi}/n_i + \sigma^2_{pj}/n_j + \sigma^2_{pij,e}/n_in_j\,.$$

For the absolute error variance the variance components in which $p$ is not involved, are relevant too. The absolute error variance is.

$$\sigma^2_{\text{Abs}} = \sigma^2_i/n_i + \sigma^2_j/n_j + \sigma^2_{ij}/n_in_j + \sigma^2_{pi}/n_i + \sigma^2_{pj}/n_j + \sigma^2_{pij,e}/n_in_j\,.$$

5.7 In this exercise we have a design in which persons are nested within judges. The variance of the mean score for a judge, for random samples of 50 persons, equals 2.0 (the person variance within judges divided by sample size, i.e. the number of persons). If there are no differences between judges, a variance between judges equal to 2.0 is expected. The variance between judges is higher: 9.0. Obviously judges are not equally lenient. One might consider the possibility to correct psychometrically for the differences in leniency. A full correction for the obtained mean differences between judges does not seem appropriate. For, part of the differences might be due to differences between the random samples of persons. The reliability of the effects of the judges equals 7.0 (observed variance for judges minus error variance) divided by 9.0 (observed variance). Application of Kelley's formula gives an estimated effect for judge 1 equal to (7.0/9.0) × (32.0 - 35.0) = -2.33. The effect for judge 2 is estimated as 0.0 and the estimated effect for judge 3 is 2.33. Judge 1 is too harsh. One might correct for this by adding 2.33 to all judgements by this judge.

The analysis is not fully satisfactory. The estimation of the true variation between judges is based on just three judges. The compromise between no correction at all for judges and a full correction is, however, attractive. One might use the outcome not only for the purpose of statistically correcting scores. The outcome might also stimulate the reorientation of the training of judges and the reformulation of judgmental instructions.

6.1 In this exercise the binomial model is to be used with parameters $\varsigma = 0.8$ and $n = 10$. The probabilities of 8, 9 and 10 correct responses have to be summed. The probability of 8 correct is 0.3020, the probability of 9 correct 0.2684 and the probability of 10 correct 0.1074. The probability of 8 or more items correct equals 0.678.

6.2 a. The proportion of correct responses for item 8 is 0.65, the item-rest correlation is 0.543. The item-test regression is given in the table below. For this particular item with a high item-rest correlation the proportion of correct responses as a function of total score increases strongly in the score range 3 – 7.

| total score | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| prop. correct | - | - | 0.0 | - | 0.33 | 0.25 | 0.50 | 1.0 | 1.0 | 1.0 | - |

b. Item 6 has the lowest item-rest correlation. The correlation is unsatisfactory (it is negative!). So, item 6 should be eliminated first. Also, when item 6 is dropped, the increase in coefficient alpha is highest.

6.3 We can estimate the reliability with KR21. The average proportion correct equals 0.75, the test variance equals 2.25. The resulting estimate equals 0.185.

6.4 Here we have an application of Equation (6.11).
$P(x_1 = 1, x_2 = 1|\varsigma) = 0.56$,
$P(x_1 = 1, x_2 = 0|\varsigma) = 0.14$,
$P(x_1 = 0, x_2 = 1|\varsigma) = 0.24$,
$P(x_1 = 0, x_2 = 0|\varsigma) = 0.06$.

6.5 Keats' solution assumes that the variance of item means given true score is relatively high in the middle score range. We can compute the item-test regressions like in Figure 6.1. When many item-test regressions cross each other in the middle score range, this assumption is untenable.

6.6 The error variance is 0.6 × 0.4 + 0.7 × 0.3 + 0.8 × 0.2 = 0.61.

The true proportion correct $\varsigma_p$ is equal to 0.7. The binomial error variance is slightly higher: 3 × 0.7 × (1 – 0.7) = 0.63.

The variance of the item difficulties at $\varsigma = \varsigma_p$ is $(0.1^2 + 0.0^2 + 0.1^2)/3 = 0.02/3$. The difference between the binomial error variance and the variance in the generalized binomial model is equal to 3 × 0.02/3.

6.7 The covariance between the item and the total test is $(n-1)\text{cov} + s^2$, where $n$ is the number of items, cov is the covariance between the items and $s^2$ the variance of the item. The variance of the test scores is $ns^2 + n(n-1)\text{cov}$. The item-total correlation can be computed from the covariance and the variances, the item-rest correlation can be computed using (6.20). The results are given in the following table:

| $n$ | $r_{it}$ | $r_{ir}$ |
|---|---|---|
| 10 | 0.529 | 0.372 |
| 20 | 0.490 | 0.406 |
| 40 | 0.469 | 0.426 |

The $r_{it}$ is spuriously high. The effect of the item as part of the total test strongly diminishes as test length increases. The value of $r_{it}$ decreases. The value $r_{ir}$ only depends on the reliability of the rest-test. The reliability increases with test length; so, $r_{ir}$ increases with test length. The difference between the two indices of item discrimination power decreases with increasing test length.

7.1 The fact that persons have been selected does not imply that the variance diminishes. In the exercise we have an example of selection which results in a higher variance in the selected group. Application of Formula (7.3) gives the value of 0.66 for the correlation in the total group.

7.2 We compare the relative frequencies of A and B for each score level. In order to obtain these relative frequencies we multiply the frequencies of group A by four

| score | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $4\times f_A$ | 0.172 | 0.436 | 0.520 | 0.696 | 0.868 | 0.696 | 0.348 | 0.172 | 0.088 | 0.0 | 0.0 |
| $f_B$ | 0.0 | 0.0 | 0.0 | 0.0 | 0.045 | 0.091 | 0.136 | 0.182 | 0.227 | 0.182 | 0.136 |

From this table the posterior probabilities can be obtained. For score level 7, for example, the posterior probability for A equals 0.172/(0.172 + 0.182) and the posterior probability for B equals 0.182/(0.172 + 0.182). From score 7 onwards the (posterior) probability of dealing with a B-person is higher than the probability of dealing with an A-person. From score 7 through score 10 we therefore classify a person as belonging to group B. With persons who belong to B we make a wrong classification in 27.2 percent of the cases (all B-persons with a score lower than 7). With persons belonging to population A we make a mistake in only 6.5 percent of the cases (all A-persons with a score equal to or higher than 7). In 1 out of the 5 cases a B-person is involved and then we make a wrong classification in 27.2 percent of the cases, in 4 out of the 5 cases we have an A-person and then we make a wrong classification in 6.5 percent of the cases (see the frequency distribution $f_A$). On the average we make a mistake in $100 \times (0.2 \times 0.272 + 0.8 \times 0.065) = 10.6$ percent of the cases. The relatively large error with respect to population B is due to the fact that this population is so much smaller than population A.

7.3 The posterior probability of belonging to group B increases for every score level, as might be inferred from Figure 7.2 and Equation 7.4. Therefore the critical score for allocation to B instead of A moves to the left. The data in the table of exercise 7.2 are relevant for a base rate equal to 0.5. We notice that the probability of belonging to B exceeds the probability of belonging to A at a score equal to or higher than 6. So, persons with a score equal to or higher than 6 can be classified as belonging to group B. Many more persons are classified as B-persons.

7.4 The optimal cut score is the score for which the expected true score equals the criterion of mastery, 0.70. The expected true score for a given score $x$ is given by Kelley's formula. With the given criterion, mean score and reliability we find that the value for the optimal cut score is 55.0 (from $70.0 = 0.25x + (1 - 0.25) \times 75$), well below the criterion on the true score scale. This is due to the low test reliability.

7.5 The proportion of correct classifications is 0.80. The proportion of correct classifications expected by chance is $0.7^2 + 0.3^2 = 0.58$. The value of $\kappa$ is 0.524.

8.1 The answers to exercise 8.1 are given in the table below.

| $\theta$ | -2.0 | -1.5 | -1.0 | -0.5 | 0.0 | 0.5 | 1.0 | 1.5 | 2.0 |
|---|---|---|---|---|---|---|---|---|---|
| $P(\theta)$ | 0.12 | 0.18 | 0.27 | 0.38 | 0.50 | 0.62 | 0.73 | 0.82 | 0.88 |

8.2 In the Rasch model one restriction is needed in order to fix the latent scale. We take the restriction $b_1 = 0$. Next, we estimate $\theta_1$ using the logarithm of $P_1(\theta_1)/[1 - P_1(\theta_1)]$ . This logarithm equals $\theta_1 - b_1 = \theta_{1.}$ The value of $\theta_1$ is -0.5. In a similar way we obtain $\theta_2$; $\theta_2$ equals 0.5.

Knowing $\theta_1$ we can compute the item parameter of the second item, $b_2 = - \ln\{ P_2(\theta_1)/[1 – P_2(\theta_1)]\} + \theta_{1.}$ The value of $b_2$ is -0.25. The value of $b_2$ can also be obtained from the equation $b_2 = - \ln\{ P_2(\theta_2)/[1 – P_2(\theta_2)]\} - \theta_{2.}$ Using this equation we obtain -0.40 as the value of $b_2$. The second computation of the item parameter is not in agreement with the first computation. This means that the Rasch model cannot describe the probabilities.

8.3 The values of the likelihood for the different levels of $\theta$ from the exercise are given in the fifth column in the table below:

| $\theta$ | $P_1(\theta)$ | $P_2(\theta)$ | $Q_3(\theta) = 1 – P_3(\theta)$ | $P(\mathbf{x}\|\theta)=P_1(\theta)P_2(\theta)Q_3(\theta)$ | $\theta P(\mathbf{x}\|\theta)g(\theta)/ P(\mathbf{x})$ |
|---|---|---|---|---|---|
| -1.0 | 0.3775 | 0.2689 | 0.8176 | 0.0830 | -0.0957 |
| -0.5 | 0.5000 | 0.3775 | 0.7311 | 0.1380 | -0.0795 |
| 0.0 | 0.6225 | 0.5000 | 0.6225 | 0.1937 | 0.0 |
| 0.5 | 0.7311 | 0.6225 | 0.5000 | 0.2275 | 0.1311 |
| 1.0 | 0.8176 | 0.7311 | 0.3775 | 0.2257 | 0.2600 |
| | | | | $P(\mathbf{x}) = 0.8679\times0.20$ | EAP = 0.2159 |

a) The maximum value of the likelihood $P(\mathbf{x}|\theta)$ is obtained in the table for $\theta = 0.5$. The ML estimate of $\theta$ must lie in the neighborhood of this value of $\theta$, i.e. between 0.0 and 1.0. For a value of $\theta$ slightly above 0.5 the likelihood exceeds the likelihood at $\theta = 0.5$ (The derivative of the log likelihood (8.38) is positive at $\theta = 0.5$). So, the ML estimate lies in the interval 0.5 – 1.0.

b) Due to the equality of the latent classes the computation of $P(\mathbf{x})$ can be simplified. The value of the EAP is 0.22. This estimate is closer to the population mean than the ML estimate.

8.4 The information values (see (8.32)) for the items are given in the rightmost column of the following table.

| item | $p(\theta)$ | $P(\theta)$ | $p(\theta)[1 - p(\theta)]$ | $I(\theta)$ |
|---|---|---|---|---|
| 1 | 0.378 | 0.378 | 0.235 | 0.235 |
| 2 | 0.269 | 0.269 | 0.197 | 0.786 |
| 3 | 0.269 | 0.452 | 0.197 | 0.351 |

8.5 The true variance equals 0.300. The error variance for each level of $\theta$ is obtained by taking $1/I(\theta)$. The average error variance equals 0.108. The reliability (true variance divided by the sum of the true variance and the average error variance) equals 0.735.

9.1 The question itself is biased. It is possible that the second item is biased against the focal group, but it is equally possible that item 1 is biased against the reference group. There is not enough evidence to choose between these two rival hypotheses and the third hypothesis of no bias. The difference between the outcomes of the two items might also be due to a higher discrimination of item 2.

9.2 We are looking for a two-item test for which the minimum of $I(\theta_1)$ and $I(\theta_2)$ is maximal. The test information equals the sum of the item informations. The item informations for the two levels of θ are:

| $b_i$ | $I_i(\theta_1)$ | $I_i(\theta_2)$ |
|---|---|---|
| -0.50 | 0.2500 | 0.1966 |
| -0.30 | 0.2475 | 0.2139 |
| 0.0 | 0.2350 | 0.2350 |
| 0.25 | 0.2179 | 0.2461 |
| 0.50 | 0.1966 | 0.2500 |

For the combination of items 2 and 4 $\min\{I(\theta_1),I(\theta_2)\}$ equals 0.46, obtained at $\theta_2$. All other item combinations have a smaller value for the minimum of $I(\theta_1)$ and $I(\theta_2)$. So, the combination of item 2 and 4 is optimal.

9.3 The next item to be presented is the item with the highest item information at $\theta = 0.20$. The item informations for this value of θ are: 0.22, 0.82, 0.12, 0.25, 0.53. The second item has the highest item information (0.82). Therefore item 2 is the item to be presented next.

10.1 From Formulas (10.2) – (10.4) we obtain $y' = 1.0625y - 8.75$. The score y= 50 is equivalent to $x$ score 44.375.

10.2 If the group that made test $Y$, is a bit better, then the average score on this test is somewhat too high in comparison to the average $x$ score. The factor $B$ from Formula (10.4) is too low, as well as the equivalent score on $X$, obtained in this way. The equivalent score $x$ should be higher than 44.375, the value obtained in Exercise 10.1. The correction that is needed can be provided by information on an anchor test $V$. The average score $v$ in the group persons to which test $Y$ was administered, is higher than the average score $v$ in the group to which test $X$ was administered. This results in a higher value for the equivalent score $x$ in Formula (10.8).

10.3 In order to obtain test $X$ subtest $V$ must be lengthened by a factor $k$. The observed-score variance of test $X$ can be written as

$$\text{(a)} \qquad \sigma_X^2 = k^2\sigma_{\mathrm{T}}^2 + k\sigma_E^2\,,$$

where $\sigma_{\mathrm{T}}^2$ is the true-score variance of subtest $V$ and $\sigma_E^2$ the error variance of subtest $V$. Let $V$ be the first of the $k$ subtests. The covariance between test $X$ and subtest $V$ is

$$\text{(b)} \qquad \sigma_{XV} = \mathrm{cov}(k\mathrm{T} + \sum_{i=1}^{k} E_i, \mathrm{T} + E_1) = k\sigma_{\mathrm{T}}^2 + \sigma_E^2\,.$$

If we divide the result of (a) by the result of (b) we obtain factor $k$.

10.4 The average $b$ of the common items is equal to 1.0 for test $Y$ and 0.5 for test $X$. In order to bring the $b$-parameters from test $Y$ to the scale of test $X$ the following transformation is to be applied: $b_Y{}^* = b_Y - 1.0$ ( - average parameter value of item 3 and 5 in $Y$) + 0.5 ( + average parameter value of item 3 and 5 in $X$): $b_Y{}^* = b_Y - 0.5$.

The item parameters of the items from test $Y$ on the scale of $X$ are: -2.0, -1.5, -0.5, 0.0, 1.5. The true scores on tests $X$ and $Y$ for a given value of θ are computed by means of Equation (6.17).

The relation between the true score for θ = -4.0 (0.5) 4.0 is:

| $\theta$ | $\tau_Y$ | $\tau_X$ |
|---|---|---|
| -4.0 | 0.25 | 0.11 |
| -3.5 | 0.39 | 0.18 |
| -3.0 | 0.59 | 0.29 |
| -2.5 | 0.86 | 0.45 |
| -2.0 | 1.21 | 0.70 |
| -1.5 | 1.62 | 1.04 |
| -1.0 | 2.08 | 1.48 |
| -0.5 | 2.55 | 2.00 |
| 0.0 | 3.00 | 2.55 |
| 0.5 | 3.43 | 3.08 |
| 1.0 | 3.80 | 3.56 |
| 1.5 | 4.12 | 3.96 |
| 2.0 | 4.38 | 4.28 |
| 2.5 | 4.58 | 4.51 |
| 3.0 | 4.72 | 4.68 |
| 3.5 | 4.82 | 4.80 |
| 4.0 | 4.89 | 4.87 |

For all true scores on *Y* the corresponding true scores on *X* are lower: test form *Y* is the easier test form.

10.5 In the first study the two examinee groups are defined on basis of an external criterion. In the second study - study with two-stage testing - the group definition is based on the common routing test *V*. Due to the fact that test *V* is not perfectly reliable, we have a regression effect on the test given as second test. Study 1 presents the design of a vertical equating study with common or anchor test *V*. In study 1 unbiased item parameter estimates for all three tests can be obtained. This is not the case in study 2 unless the regression effect is effectively dealt with. Application of the mean and sigma method would not result in correct estimates on a common scale.

# REFERENCES

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Akkermans, W. (2000). Modelling sequentially scored item responses. *British Journal of Mathematical and Statistical Psychology*, *53*, 83-98.

Alf, E. F., & Dorfman, D. D. (1967). The classification of individuals into two criterion groups on the basis of a discontinuous payoff function. *Psychometrika, 32*, 115-123.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*, 185-198.

American Council on Education (1995*). Guidelines for computer-adaptive test development and use in education*. Washington, DC.

American Educational Research Association (1955). *Technical recommendations for achievement tests*. Washington, DC: National Education Association.

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51 (Suppl.).

American Psychological Association (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1974). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1954). *Psychological testing*. New York: MacMillan.

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34*, 42-54.

Andersen, E. B. (1973). A goodness of fit test for the Rasch-model. *Psychometrika, 38*, 123-40.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69-81.

Andersen, E. B. (1983). A general latent structure model for contingency data. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement*. Lawrence Erlbaum: N.J, pp. 117-38.

Anderson, N. H. (1961). Scales and statistics: parametric and nonparametric. *Psychological Bulletin*, 58, 305-316.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Andrich, D. (1999). Rating scale analysis. In J. P. Keeves & G. N. Masters (Eds.), *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon, pp. 110-21.

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika*, 18, 1-14.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington: American Council on Education, 509-600.

Assessment Systems Corporation. (1996*). User's manual for the XCALIBRE marginal maximum-likelihood estimation program*. St. Paul, MN: Assessment Systems Corp.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Bickel, P., Buyske, S., Chang, H., & Ying, Z. (2001). On maximizing item information and matching difficulty with ability. *Psychometrika, 66*, 69-77.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Blinkhorn, S. F. (1997). Past imperfect, future conditional: Fifty years of test theory. *British Journal of Mathematical and Statistical Psychology*, *50*, 175-185.

Blok, H. (1985). Estimating the reliability, validity and invalidity of essay ratings. *Journal of Educational Measurement*, *22*, 41-52.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46*, 443-459.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 13*, 261-280.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Braun, H. I. (1988). Understanding scoring reliability: experiments in calibrating essay readers. *Journal of Educational Statistics, 13*, 1-18.

Brennan, R. L. (1987). Problems, perspectives, and practical issues in equating. *Applied Psychological Measurement*, *11*, 221-306.

Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: American Testing Program (rev. ed.).

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, *22*, 307-331.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277-289.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education, pp. 367-407.

Burr, J. A., & Neselroade, J. R. (1990). Change measurement. In A. von Eye (Ed.), *Statistical methods in longitudinal research*. Vol 1. Boston: Academic Press, pp. 3-34.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and divergent validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Carlson, J. F. (1998). Review of the Beck Depression Inventory (1993 Revised). In Impara, J. C., & B. S. Plake (Eds.), *The thirteenth mental measurement yearbook*. Incoln, NE: The Buros Institute of Mental Measurements, pp. 117-19.

Cizek, G. J. (2001). (Ed.*) Setting performance standards: concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, A. S., & Kim, S.-H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement*, *22*, 116-130.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch-model. *British Journal of Mathematical and Statistical Psychology, 32*, 113-120.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education, pp. 201-19.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*, 225-244.

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi experimentation. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press, pp. 491-576.

Cook, T. D., & Shadish, W. R. (1994). Social experiments & some developments over the past fifteen years. *Annual Review of Psychology*, *45*, 545-580.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Cooper, H., & Hedges, L. V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

Cowell, W. R. (1982). Item-response-theory pre-equating in the TOEFL testing program. In P.W. Holland & D. B. Rubin (Eds.), *Test equating*. New York: Academic Press, pp. 149-61.

Crano, W. D. (2000). The multitrait-multimethod matrix as synopsis and recapitulation of Campbell's views on the proper conduct of social inquiry. In L. Bickman (Ed.). *Research design: Donald Campbell's legacy* (*Vol*. 2). Thousand Oaks, CA: Sage, pp. 37-61.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.). *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 443-507.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change" or should we? *Psychological Bulletin, 74*, 68-80.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972*). The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessment of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373-399.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137-163.

Cronbach, L. J., & Warrington, W. G. (1952). Efficiency of multiple-choice tests as a function of spread of item difficulties. *Psychometrika, 17*, 127-147.

Croon, M. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44*, 315-331.

Daniel, M. H. (1999). Behind the scenes: using new measurement methods on the DAS and KAIT. In S. E. Embretson & S. L. Herschberger (Eds.)*, The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 37-63.

De Gruijter, D. N. M. (1985). Compromise models for establishing examination standards. *Journal of Educational Measurement*, *22*, 263-269.

De Gruijter, D. N. M. (1988). Standard errors of item parameter estimates in incomplete designs. *Applied Psychological Measurement, 12,* 109-116.

De Gruijter, D. N. M., & Van der Kamp, L. J. Th. (1991). Generalizability theory. In R. K. Hambleton, & J.N. Zaal (Eds.), *Advances in educational and psychological testing*. Boston: Kluwer, pp. 45-68.

De Leeuw, J. & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, 11, 183-196.

Donoghue, R. R., & Isham, P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*, 33-51.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education*, *3*, 1-17.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum, pp. 35-66.

Dorans, N. J., & Kulick, E. M. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355-368.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist*, *16*, 640-647.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175-186.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika,. 56*, 495-516.

Embretson, S. E. & Prenovost, L. K. (1999). Item response theory in assessment research. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology*. New York, Chichester: Wiley, pp. 276-94.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ercika, K., Schwartz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*, 137-154.

Fhanér, S. (1974). Item sampling and decision making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, *27*, 172-175.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement*, *44*, 883-891.

Feldt, L. S., & Brennan, R. L. S. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education, pp. 105-46.

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9*, 351-361.

Feldt, L. S., & Qualls, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement*, *33*, 141-156.

Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. de Gruijter & L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational measurement*, Wiley: New York, pp. 97-110.

Fischer, G. H. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

Fraser, C. (1988). NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, NSW: University of New England.

Furneaux, W. D. (1960). Intellectual abilities and problem-solving behaviour. In H. J. Eysenck (Ed.), *Handbook of abnormal psychology*. London: Pitman, pp. 167-92.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.

Glas, C. A. W., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, *42*, 139-167.

Gregory, R. J. (2000). *Psychological testing: History, principles, and applications* (3rd ed.). Boston: Allyn and Bacon.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Guttman, L. (1945). A basis for test-retest reliability. *Psychometrika, 10*, 255-282.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.). *Measurement and prediction*, *Vol.* IV. Princeton, NJ: Princeton University Press, pp. 60-90.

Guttman, L. (1953). A special review of Harold Gulliksen, Theory of mental tests, *Psychometrika*, *18*, 123-130.

Haebara, T. (1980). Equating logistic ability scales by weighted least squares method. *Japanese Psychological Research, 22*, 144-149.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Mahwah, NJ: Erlbaum.

Hambleton, R. K. (1996). Advances in assessment models, methods, and practices. In D. C. Berliner and R. C. Calfee (Eds.), *Handbook of educational psychology*. New York: MacMillan, pp. 899-925.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*, 159-170.

Hambleton,R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hand, D. J. (1997). *Construction and assessment of classification rules.* New York: Wiley.

Hanson, B. A., Zeng, L., & Kolen, M. J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17*, 225-237.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*, 139-164.

Hattie, J., Jaeger, R. M., & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education*, *24*, 393-446.

Heinen, T. (1996). *Latent class and discrete latent trait models. Similarities and differences*. Thousand Oaks, CA: Sage.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. *Psychometrika*, *28*, 211-218.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*, 577-601.

Holland, P. W., & Wightman, L. E. (1982). Section pre-equating: a preliminary investigation. In P. W. Holland & D. B. Rubin (Eds.), *Test equating*. New York: Academic Press, pp. 271-97.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*, 153-160.

Hunter, J. E., & Schmidt, F. L. (1990) *Methods of meta-analysis*. Newbury Park: Sage.

Huynh, H. (1978). Reliability of multiple classifications. *Psychometrika*, *43*, 317-325.

Huynh, H. (1994). On the equivalence between a partial credit model and a set of independent Rasch binary items. *Psychometrika*, *59*, 111-119.

Jackson, P. H. (1973). The estimation of true score variance and error variance in the classical test theory model*. Psychometrika, 38*, 183-201.

Jacoby, W. G. (1991). *Data theory and dimensional analysis*. Newbury Park, CA: Sage.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed*.), Educational measurement* (3rd ed.). New York: American Council on Education, pp. 485-514.

Jansen, P. G. W., & Roskam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika*, *51*, 69-91.

Jarjoura, D. (1983). Best linear prediction of composite universe scores. *Psychometrika, 48*, 525-539.

Jarjoura, D. (1986). An estimator of examinee-level measurement error variance that considers test form difficulty adjustments. *Applied Psychological Measurement*, *10*, 175-186.

Jarjoura, D., & Brennan, R. L. (1982). A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement, 6*, 161-171.

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

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce & P. Suppes (Eds.), *Contemporary developments in mathematical psychology. Vol. II.* San Francisco: Freeman, pp. 1-56.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8 user's reference guide*. Chicago: Scientific Software International.

Judd, C. M., & Kenny, D. A. (1981). *Estimating the effects of social interventions*. Cambridge: Cambridge University Press.

Judd, C. M., & McClelland, G. H. (1998). Measurement. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology*. Vol. 1 ($4^{th}$ ed.) Boston, MA: McGraw-Hill, pp. 180-232.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika, 22*, 29-41.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49*, 223-245.

Kelley, T. L. (1947). *Fundamentals of statistics*, Cambridge: Harvard University Press.

Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics*, vol. II. London: Griffin.

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*, 345-355.

Kish, L. (1987). *Statistical design for research*. Wiley: New York.

Kok, F. G., Mellenbergh, G. J., & Van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22*, 295-303.

Kolen, M. J. (1999). Equating of tests. In G. M. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon, pp. 164-75.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.

Lee, W.-C., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: a simulation study. *Journal of Educational Measurement*, *37*, 1 –20.

Levy, P. (1973). On the relationship between test theory and psychology. In P. Kline (Ed.), *New approaches in psychological measurement*. London: Wiley, p. 1-42.

Lindley, D. V. (1971). The estimation of many parameters. In V. P. Godambe and D. A. Sprott (Eds.), *Foundation of statistical inference*. Toronto: Holt, Rinehart and Winston, pp. 435-47.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement*, *18*, 109-118.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

Little, R. J. A, & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York.

Longford, N. T. (1994). Reliability of essay rating and score adjustment. *Journal of Educational and Behavioral Statistics, 19*, 171-200.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, No 7. Chicago: University of Chicago Press.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750-751.

Lord, F. M. (1954). Further comment on 'football numbers'. *The American Psychologist, 9*, 264-265.

Lord, F. M. (1955). Equating test scores – a maximum likelihood solution. *Psychometrika, 20*, 193-200.

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika, 23*, 291-296.

Lord, F. M. (1962). Cutting scores and errors of measurement. *Psychometrika*, *27*, 19-30.

Lord, F. M. (1977). Optimal number of choices per item - a comparison of four approaches. *Journal of Educational Measurement*, *14*, 33-38.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

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

Marcoulides, G. A. (1999). Generalizability theory: Picking up where the Rasch IRT model leaves off? In S. E. Embretson and S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Erlbaum, pp. 129-152.

Masters, G. N. (1982). A Rasch-model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G. N. (1999). Partial credit model. In J. P. Keeves & G. N. Masters (Eds.), *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon, pp. 98-109.

Maxwell, A. E., & Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology*, *21*, 105-116.

McDonald, R. P. (1968). A unified treatment of the weighting problem. *Psychometrika*, *33*, 351-381.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100-117.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer-Verlag, pp. 258-69.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Meredith, W., & Kearns, J. (1973). Empirical Bayes point estimates of latent trait scores without knowledge of the trait distribution. *Psychometrika*, *38*, 533-554.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education, pp. 13-103.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13-23.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*, 741-749.

Michell, J. (1999*). Measurement in psychology: Critical history of a methodological concept*. Cambridge: Cambridge University Press.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias, *Applied Psychological Measurement, 17*, 297-334.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Mokken, R. J. (1971). *A theory and procedure of scale analysis with applications in political research.* New York-Berlin: Walter de Gruyter (Mouton).

Mokken, R. J., Lewis, C., & Sijtsma, K. (1986). Rejoinder to 'The Mokken scale: a critical discussion'. *Applied Psychological Measurement, 10*, 279-285.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch-model. *Psychometrika, 48*, 49-72.

Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer-Verlag, pp. 369-380.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person-fit indices. *Psychometrika, 55*, 75-106.

Molenaar, I. W., & Sijtsma, K. (2002*). Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Mosier, C. I. (1947). A critical examination of the concept of face validity. *Educational and Psychological Measurement*, *7*, 191-205.

Mulaik, S. A. (1972). *The foundations of factor analysis*. New York: McGraw-Hill.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59-71.

Muraki, E., & Bock, R. D. (1997). PARSCALE. IRT based test scoring and item analysis for graded open-ended exercises and performance tasks. Chicago: Scientific Software.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73-90.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications*. (5th Ed.) Upper Saddle River, NJ: Prentice Hall.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Nandakumar, R., Yu, F., Li, H.-H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement*, *22*, 99-115.

NDT (2002). *Netherlands Differentiation Test: Technical Manual*. In preparation.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, *14*, 3-19.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis.* Hillsdale, NJ: Lawrence Erlbaum.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Oud, J. H. L., Van den Bercken, J. H., & Essers, R. J. (1990). Longitudinal factor score estimation using the Kalman filter. *Applied Psychological Measurement, 14*, 395-418.

Overall, J. E. (1965). Reliability of composite ratings. *Educational and Psychological Measurement, 25*, 1011-1022.

Pandey, T. N., & Hubert, L.(1975). An empirical comparison of several interval estimation procedures for coefficient alpha. *Psychometrika, 40*, 169-181.

Panter, A. T., Kimberly, A. S., & Dahlstrom, W. G. (1997). Factor analytic approaches to personality item-level data. *Journal of Personality Assessment*, *68*, 561-589.

Pearson, E. S., & Hartley, H. O. *Biometrika tables for statisticians. Vol. I*. Cambridge: University Press, 1970 (3rd ed.).

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: a comparative study of scale stability. *Journal of Educational Statistics, 8*, 137-156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education, pp. 221-62.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3-29.

Popham, W. J., & Husek, T. R. (1969) Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1-9.

Rajaratnam, N. (1960). Reliability formulas for independent decision data when reliability data are matched. *Psychometrika, 25*, 261-271.

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified parallel tests. *Psychometrika, 30*, 39-56.

Raykov, T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement*, *22*, 369-374.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25-36.

Reise, S. P. (1999). Personality measurement; issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Herschberger (Eds.), *The new rules of measurement: What every psychologist and educator should know*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 219-41.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*, 45-58.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552-566.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311-327.

Rogers, W. T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement*, *59*, 234-247.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 90*, 726-748.

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations*.* *Journal of Educational Statistics*, *16*, 157-252.

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton, (Eds.), *Handbook of modern item response theory*. New York: Springer-Verlag, pp. 187-208.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75-92.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, *35*, 1-30.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, *20*, 207-219.

Rudner, L. M., Getson, P. R., & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213-233.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, No 18*. Iowa City, IA: Psychometric Society.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38*, 221-233.

Samejima, F. (1979). A new family of models for the multiple-choice item. *Research Report 79-4*. University of Tennessee, Knoxville, TN.

Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.). (1997*). Computerized adaptive testing: from inquiry to operation*. Washington, DC: American Psychological Association.

Scheuneman, J. D., & Bleistein, C. A. (1999). Item bias. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon, pp. 220-34.

Schmitt, N., Coyle, B. W., & Saari, B. B. (1977). A review and critique of analyses of multitrait-multimethod matrices. *Mutivariate Behavioral Research, 12*, 447-478.

Schmitt N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement, 10*, 1-22.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory; a primer*. Newbury Park, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, *44*, 922-932.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159-194.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*, 77-105.

Sheridan, B., Andrich, D., & Luo, G. (1996). *Welcome to RUMM: A windows based item analysis program employing Rasch unidimensional measurement models*. User's Guide.

Shi, J. Q., & Lee, S. Y. (1997). A Bayesian estimation of factor score in confirmatory factor model with polytomous, censored or truncated data. *Psychometrika, 62*, 29-50.

Sirotnik, K,., & Wellington, R. (1977). Incidence sampling: An integrated theory for "matrix sampling". *Journal of Educational Measurement*, *14*, 343-399.

Skaggs, S. G., & Lissitz, R. W. (1986) IRT equating: Relevant issues and a review of recent research. *Review of Educational Research*, 56, 495-529.

Snijders, T. A. B., & Bosker, R. J. (1999*). Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: Sage.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, pp. 1-49.

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, 161, 849-856.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement,7*, 201-210.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmastery classifications. In R. A. Berk (Ed.), A guide to criterion-referenced test construction. Baltimore: Johns Hopkins University Press, pp. 267-91.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Swaminathan, H, & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589-601.

Tate, R. L. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement*, *36*, 336-346.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23*, 565-578.

Tellegen, A. (1982). *A brief manual for the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.

Ten Berge, J. M. F., Snijders, T. A. B., & Zeegers, F. E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis, *Psychometrika, 46*, 201-213.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50*, 411-420.

Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software Int.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds). *Test validity*. Hillsdale, NJ: Erlbaum, pp. 147-69.

Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology*, *26*, 249-269.

Urry, V. W. (1974). Approximation to item parameters of mental test models and their use. *Educational and Psychological Measurement, 34*, 253-269.

Van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, 22, 195-211.

Van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximum model for test design with practical constraints. *Psychometrika, 54*, 237-247.

Van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic Publishers.

Van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Van der Linden, W. J., & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement, 1*, 593-599.

Van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*, 259-270.

Van der Rijt, B. A. M., Van Luit, J. E. H., & Pennings, A. H. (1999). The construction of the Utrecht Early Mathematical Competence Scales. *Educational and Psychological Measurement*, *59*, 289-309.

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications*. New York: Springer, pp. 215-37.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, N.J.: Lawrence Erlbaum.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203-220.

Waller, N. G. (1998). Review of the Beck Depression Inventory (1993 Revised). In Impara, J. C., & B. S. Plake (Eds.), *The thirteenth mental measurement yearbook*. Incoln, NE: The Buros Institute of Mental Measurements, pp. 120-21.

Walsh, W. B., & Betz, N. E. (2001).Test and assessment. (4th Ed.) Upper saddle River, NJ: Prentice Hall.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Webb, N. M., Shavelson, R. J., Kim, K. S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. *Military Psychology*, *1*, 91-110.

Weiss, D.J., & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Werts, C. E., Breland, H. M., Grandy, J., & Rock, D. R. (1980). Using longitudinal data to estimate reliability in the presence of correlated measurement models. *Educational and Psychological Measurement*, *40*, 19-29.

Wilcox, R. R. (1976). A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, *1*, 359-364.

Wilcox, R. R. (1981). A closed sequential procedure for comparing the binomial distribution to a standard. *British Journal of Mathematical and Statistical Psychology*, *34*, 238-242.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Mooresville, IN: Scientific Software.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*, 144-152.

Woodruff, D. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement*, *27*, 191-208.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago: Mesa Press.

Wright, B. (1988). The efficacy of unconditional maximum likelihood bias correction: Comment on Jansen, van den Wollenberg, and Wierda. *Applied Psychological Measurement*, *12*, 315-318.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago: Scientific Software.

Zwick, R. (1990). When do item response function and Mantel-Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15*, 185-197.

# AUTHOR INDEX

# SUBJECT INDEX