Charles University in Prague
Faculty of Mathematics and Physics

Department of Probability and Mathematical Statistics

DOCTORAL THESIS

Reliability of measurements consisting of dichotomously scored items

Patrícia Martinková

Advisor: doc. RNDr. Karel Zvára, CSc.

Prague 2007

I hereby declare that I wrote the thesis on my own and that I have included all the sources of information which I exploited to the references. I authorize Charles University to lend this document to other institutions or individuals for academic and research purposes.

Prague, April 11, 2007 Patrícia Martinková

To my children

Acknowledgements

I am particularly grateful to my advisor, doc. RNDr. Karel Zvára, CSc., for his patient guidance over the years of my study, for the countless hours of discussion, insightful advice and for a careful proofreading of this thesis.

I would also like to thank the many people who discussed with me various topics related to this work and who proofread components of the thesis. Many thanks belong to all professors and students of the Department of Probability and Mathematical Statistics and to all my colleagues from the EuroMISE Center for the friendly and stimulative environment.

Most importantly, I wish to thank my family for helping me in the tough times of finishing my thesis: my parents, grandmother and mother in law for taking care of my son, and Šimůnek for being a good boy and keeping me in a positive and creative mood. My deepest thanks belong to my husband Igor for his endless caring support.

Finally, I wish to acknowledge the financial support of the grant 1M06014 of the Ministry of Education of the Czech Republic.

Contents

Preface

The thesis deals with reliability of measurements in the context of multiple-item testing instruments, such as educational tests. It starts from author's diploma thesis, where the reliability was studied in the context of the classical test theory (that is under the assumptions of ANOVA models). Some of the basic definitions and properties appear here as Chapter 1. The rest of the thesis is dedicated to the measurements composed of dichotomous items, for which the classical model is not appropriate.

While there are numerous papers written on reliability in the context of the classical test theory (see for example Hoyt (1941), Guttman (1945), Cronbach (1951), Novick and Lewis (1967), etc.), not so much has been written about reliability in the case when the assumptions of ANOVA models are not fulfilled. Some works discuss reliability in the presence of outliers and propose robust estimators of reliability (see Wilcox (1992) and more recently Christmann and Van Aelst (2006)). Even though there is some work devoted to reliability of binary data (see for example Ridout et al. (1999), Zou and Donner (2004)), it is mostly based on common correlation models, where reliability merges with intraclass correlation coefficient. Unfortunately, such models (for example the beta-binomial model) are not appropriate for multiple-item testing instruments, since they do not allow for different item difficulties.

The present thesis provides an extension of the reliability concept. A more general definition of reliability is proposed, of which the classical definition is shown to be a special case. The new definition is applied to models appropriate for dichotomous items. A new estimate of reliability of composite dichotomous measurements is also introduced, an estimate which in some situations seems to have better properties than the conventional estimate based on Cronbach's alpha.

A short overview of the thesis chapters follows:

Chapter 1 contains a summary of definitions and basic properties of the reliability in context of the classical test theory. The classical definition of reliability of measurement is discussed. The characteristic used most often to estimate reliability (Cronbach's alpha) and its relationship with reliability is investigated. Finally, reliability and Cronbach's alpha are studied in the framework of ANOVA models.

In Chapter 2, the models appropriate for measurements composed of dichotomously scored items are studied. In Section 2.1, the model used for dichotomously scored items most often (the Rasch model) is discussed. The techniques of parameter estimation based on the method of maximum likelihood are described. As a new result, it is shown that the conventional estimator of item difficulty based on the least squares method in the ANOVA model can be understood as an approximation of the ML estimators based on the Rasch model. The rest of Chapter 2 is dedicated to the *modified beta-binomial model*, and to its extension which we call the *general model with additive item effect* .

The general definition of reliability proposed in this work was motivated by a paper of Commenges and Jacqmin (1994), where some possibly equivalent definitions of common correlation models are discussed. In Chapter 3, we revise the work of Commenges and Jacqmin, find counterexamples to some of their results and postulate corrected propositions.

In Chapter 4, we propose a more general definition of reliability and we show the legitimacy of the new definition. Also, we derive the formulas for reliability in the models discussed in the previous chapters.

Finally in Chapter 5, the new estimate of reliability of composite dichotomous measurements is introduced and studied via simulations in the Rasch model and in the modified Rasch model.

The core of the thesis is based on author's publications:

[1] Rexová, P. (2003). *Reliability of measurements [Spolehlivost měření, in Czech]. Diploma thesis.* Charles University in Prague.

[2] Rexová, P. (2004). Item analysis of educational tests. In Šafránková, J., editor, *WDS'04 Proceedings of Contributed Papers: Part I - Mathematics and Computer Sciences*, pages 77–83. Matfyzpress, Prague.

[3] Martinková, P., Zvára, K., Zvárová, J., and Zvára, K. (2006). The new features of the ExaMe evaluation system and reliability of its fixed tests. *Methods of Information in Medicine*, 45:310–315.

[4] Martinková, P. (2006). Reliability in the Rasch model. In Hakl, F., editor, *Proceedings of the XI. PhD. Conference of the ICS ASCR*, pages 64–71. Matfyzpress, Prague.

[5] Martinková, P. and Zvára, K. (2007). Reliability in the Rasch model. *Kybernetika.* Submitted.

Chapter 1 summarizes author's diploma thesis [1]. Some results discussed here together with practical application were published in [3]. Section 2.1 was presented in [2]. Parts of Chapters 2, 4 and 5 discussing the modified beta-binomial model were published in [4]. Parts of Chapters 4 and 5 discussing the reliability in the Rasch model are contained in [5]. Chapter 3 is being prepared for publication.

Chapter 1

Introduction to reliability

1.1 Classical definition of reliability

When describing reliability of measurement, it is usually assumed that *the measurement Y* is composed of two random variables: an unobservable *true value T* and an *error term e*,

$$Y = T + e. \tag{1.1}$$

The error term is supposed to have a zero mean $\mathrm{E}\,(e) = 0$, a positive variance, and to be independent of the true value T. Therefore we have

$$\mathrm{var}\,(Y) = \mathrm{var}\,(T) + \mathrm{var}\,(e).$$

The *reliability* of such a measurement is defined by the ratio

$$R = \frac{\mathrm{var}\,(T)}{\mathrm{var}\,(Y)} = 1 - \frac{\mathrm{var}\,(e)}{\mathrm{var}\,(Y)} \tag{1.2}$$

and it compares the variance of the error term with the variance of the measured property. The smaller is the error variance relative to the observed score variance, the more reliable is the measurement. Thus, the measurement is considered to be reliable when the value of reliability is close to 1.

It should be pointed out that reliability is sample-dependent: a certain educational test can have a different reliability when given to a population with a high variability of tested knowledge and when given to a population with a low variability of the knowledge.

The following simple lemmas give a natural interpretation of the reliability.

Lemma 1.1. *Having two independent measurements $Y_1 = T + e_1, Y_2 = T + e_2$ of the same property T, where $\mathrm{var}\,(e_1) = \mathrm{var}\,(e_2)$, the reliability can be expressed as the correlation between these two measurements, $R = \mathrm{corr}(Y_1, Y_2)$.*

Proof:

$$\mathrm{corr}(T + e_1, T + e_2) = \frac{\mathrm{cov}(T + e_1, T + e_2)}{\sqrt{\mathrm{var}^2(Y_1)}} = \frac{\mathrm{cov}(T, T) + 0}{\mathrm{var}\,(Y_1)} = \frac{\mathrm{var}\,(T)}{\mathrm{var}\,(Y_1)} = R.$$

□

In terms of educational tests, the reliability reflects to what extent a test gives the same result when taken repeatedly by the same person under the same conditions.

Lemma 1.2. *The reliability can be expressed as the squared value of the correlation between the observed score and the true score,* $\operatorname{corr}^2(Y, T)$.

Proof:

$$\operatorname{corr}^2(Y,T) = \frac{\operatorname{cov}^2(T+e,T)}{\operatorname{var}(Y)\operatorname{var}(T)} = \frac{\operatorname{var}^2(T)}{\operatorname{var}(Y)\operatorname{var}(T)} = \frac{\operatorname{var}(T)}{\operatorname{var}(Y)} = R.$$

□

Thus, the reliability of an educational test measures the strength of the relationship between the score reached by a student and his/her true knowledge.

Unfortunately, none of these representations is useful when estimating the reliability of educational tests because they cannot be directly estimated from the observed data. We cannot estimate the error variance $\operatorname{var}(e)$, the true score T, nor the knowledge of a student by the same test twice and independently. Therefore, when estimating the reliability of an educational test, we must take into account the fact that the test is a composite measurement.

1.2 Reliability of a composite measurement

Let us consider the problem of measuring the reliability of a multiple-item testing instrument, such as an educational test. Consider a series of items $Y_j = T_j + e_j$, for $j = 1, \ldots, m$, where the error terms e_j are mutually independent and independent of the true scores T_k for $k = 1, \ldots, m$, having the same variance $\operatorname{var}(e_j) = \sigma_e^2 > 0$, and mean $\operatorname{E} e_j = 0$. The observed overall score of the m items is given by $Y = Y_1 + \cdots + Y_m$ and the unobservable overall true score is given by $T = T_1 + \cdots + T_m$. The reliability of such a *composite measurement* is defined by (1.2) and with regard to the above mentioned assumptions it can further be expressed as

$$R_m = \frac{\operatorname{var}(T)}{\operatorname{var}(Y)} = \frac{\operatorname{var}(T)}{\operatorname{var}(T) + \operatorname{var}(\sum e_j)} = \frac{\operatorname{var}(T)}{\operatorname{var}(T) + m\sigma_e^2}. \tag{1.3}$$

To study the relationship between the reliability of a composite measurement and the reliability of an item, let us define the *essential τ-equivalence* of items.

Definition 1.3. Items $j = 1, 2, \ldots, m$ are said to be *essentially τ-equivalent* if there exist constants $c_1, c_2, \ldots, c_m$ (we can require $\sum_j c_j = 0$) and a random variable T such that with probability equal to one it holds that

$$T_j = T + c_j. \tag{1.4}$$

Lemma 1.4. *The m items are essentially τ-equivalent if and only if for the items' true score the following holds simultaneously*

$$\operatorname{var}(T_1) = \cdots = \operatorname{var}(T_m) = \sigma_T^2, \tag{1.5}$$

$$\operatorname{corr}(T_j, T_k) = 1, \qquad j, k = 1, \ldots, m. \tag{1.6}$$

Proof: The first implication is obvious: If

$$T_j = T + c_j \qquad j = 1, \ldots, m,$$

then also

$$\operatorname{var}(T_1) = \cdots = \operatorname{var}(T_m) = \operatorname{var} T,$$

$$\operatorname{corr}(T_j, T_k) = \frac{\operatorname{cov}(T, T)}{\sqrt{\operatorname{var} T}\sqrt{\operatorname{var} T}} = 1, \qquad j, k = 1, \ldots, m.$$

Vice-versa, if (1.5) and (1.6) hold, then the conditions for equality in Cauchy-Schwarz inequality must hold. Therefore for all $j, k = 1, \ldots, m$ with probability equal to 1, the couples T_j and T_k are related by

$$T_j = T_k + c_{jk}. \tag{1.7}$$

When summing (1.7) over k and setting $T = \frac{1}{m}\sum_k T_k$, $c_j = \frac{1}{m}\sum_k c_{jk}$ we finally obtain (1.4). Another summing over j can show that constants c_j defined in this way have zero sum. □

In the following lemma, the **Spearman-Brown formula** (1.8) gives a relationship between the reliability of a composite measurement R_m and the reliability of an item R_1 for measurements composed of essentially τ-equivalent items.

Lemma 1.5. *For a measurement composed of essentially τ-equivalent items all the reliabilities R_1 of the items are equal and the reliability of the whole test can be expressed as*

$$R_m = \frac{mR_1}{1 + (m-1)R_1}. \tag{1.8}$$

Proof: From (1.4) follows

$$\operatorname{var}\left(\sum_{j=1}^{m} T_j\right) = \operatorname{var}(mT) = m^2\sigma_T^2.$$

Further, we have

$$\operatorname{var}\left(\sum_{j=1}^{m} Y_j\right) = \operatorname{var}(mT) + \operatorname{var}\left(\sum_{j=1}^{m} e_j\right) = m^2\sigma_T^2 + m\sigma_e^2.$$

Therefore

$$\begin{aligned} R_m &= \frac{\operatorname{var}\left(\sum_{j=1}^{m} T_j\right)}{\operatorname{var}\left(\sum_{j=1}^{m} Y_j\right)} = \frac{m^2\sigma_T^2}{m^2\sigma_T^2 + m\sigma_e^2} = \frac{m\frac{\sigma_T^2}{\sigma_T^2+\sigma_e^2}}{1 + (m-1)\frac{\sigma_T^2}{\sigma_T^2+\sigma_e^2}} \\ &= \frac{mR_1}{1+(m-1)R_1}. \end{aligned}$$

□

More generally:

Lemma 1.6. *Let a measurement be composed of m_1 essentially τ-equivalent items. Let the reliability of this measurement be R_{m_1}. Let us assign R_{m_2} the reliability of a measurement composed of m_2 items, so that all the items of the two measurements are essentially τ-equivalent. Then the relationship between the reliabilities R_{m_1} and $R_{m_2}$is given by*

$$R_{m_2} = \frac{\frac{m_2}{m_1} R_{m_1}}{1 + (\frac{m_2}{m_1} - 1) R_{m_1}}. \tag{1.9}$$

Proof: Lemma 1.5 implies for both composite measurements $i = 1, 2$ that

$$R_{m_i} = \frac{m_i R_1}{1 + (m_i - 1) R_1}.$$

Therefore we have

$$R_1 = \frac{\frac{1}{m_1} R_{m_1}}{1 + (\frac{1}{m_1} - 1) R_{m_1}} = \frac{\frac{1}{m_2} R_{m_2}}{1 + (\frac{1}{m_2} - 1) R_{m_2}},$$

which immediately implies (1.9). □

The consequence of this lemma is the fact that the reliability of an educational test is dependent on the number of its items. Therefore, by adding suitable items to the test, the reliability could approach as close to 1 as we would desire. When comparing reliabilities of two educational tests, which in principle cannot have the same number of items, we should bear this property of reliability in mind.

1.3 Cronbach's alpha – estimator of reliability

A widely used characteristic of reliability is called Cronbach's alpha. It was proposed as a generalization of Kuder-Richardson formula 20 for binary data (see Kuder and Richardson (1937)) and it was deeply studied in Cronbach (1951). Cronbach's alpha is defined as

$$\alpha_{CR} = \frac{m}{m-1} \frac{\mathrm{var}\,(Y) - \sum_j \mathrm{var}\,(Y_j)}{\mathrm{var}\,(Y)} = \frac{m}{m-1} \frac{\sum\sum_{j \neq k} \mathrm{cov}(Y_j, Y_k)}{\sum\sum_{j,k} \mathrm{cov}(Y_j, Y_k)}. \tag{1.10}$$

A pleasant property of Cronbach's alpha is the fact that this characteristic is easy to estimate from the data simply by using sample variances s_{jj} and sample covariances s_{jk} instead of their population counterparts in (1.10):

$$\hat{\alpha}_{CR} = \frac{m}{m-1} \frac{\sum\sum_{j \neq k} s_{jk}}{\sum\sum_{j,k} s_{jk}}, \tag{1.11}$$

where

$$s_{jk} = \frac{1}{n-1} \sum_{t=1}^{n} (Y_{tj} - \bar{Y}_{\bullet j})(Y_{tk} - \bar{Y}_{\bullet k}). \tag{1.12}$$

In this notation, the bullet stands for the sum over the replaced index and the bar over Y stands for the average:

$$\bar{Y}_{\bullet j} = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}. \tag{1.13}$$

As a consequence of the mentioned pleasant property, Cronbach's alpha is widely used for the estimation of reliability of composite measurement. Moreover, Cronbach's alpha is often mistaken for the reliability itself. Nevertheless, the equality of Cronbach's alpha and reliability holds only in some special cases as stated in the following theorem (see also Novick and Lewis (1967)):

Theorem 1.7 (Novick & Lewis). *Let $Y = Y_1 + \cdots + Y_m$ be a composite measurement with true scores $T = T_1 + \cdots + T_m$. Then Cronbach's alpha α_{CR} is the lower bound of the reliability:*

$$R_m = \frac{\operatorname{var}(T)}{\operatorname{var}(Y)} \geq \frac{m}{m-1}\left(1 - \frac{\sum_j \operatorname{var}(Y_j)}{\operatorname{var}(Y)}\right) = \alpha_{CR} \tag{1.14}$$

with equality holding if and only if the measurements $Y_1, \ldots, Y_m$ are essentially τ-equivalent.

Proof: Let us assign $\operatorname{cov}(T_i, T_j) = \sigma_{ij}$ and start with the inequality

$$0 \leq (\sqrt{\sigma_{ii}} - \sqrt{\sigma_{jj}})^2 = \sigma_{ii} + \sigma_{jj} - 2\sqrt{\sigma_{ii}\sigma_{jj}}, \tag{1.15}$$

therefore:

$$\sigma_{ii} + \sigma_{jj} \geq 2\sqrt{\sigma_{ii}\sigma_{jj}}.$$

The Cauchy-Schwarz inequality gives

$$\sqrt{\sigma_{ii}\sigma_{jj}} \geq |\sigma_{ij}| \geq \sigma_{ij}, \tag{1.16}$$

(in other words the upper bound of the correlation coefficient $\rho_{ij} = \operatorname{corr}(T_i, T_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ is one). Together we have

$$\sigma_{ii} + \sigma_{jj} \geq 2\sigma_{ij}.$$

Summing over all $i \neq j$ and dividing by $2(m-1)$ gives

$$\sum_j \sigma_{jj} \geq \frac{1}{m-1}\sum\sum_{i \neq j} \sigma_{ij}.$$

Since

$$\sum_i \sum_j \sigma_{ij} = \sum_j \sigma_{jj} + \sum\sum_{i \neq j} \sigma_{ij},$$

the last inequality can be written as

$$\sum_i \sum_j \sigma_{ij} \geq \frac{m}{m-1}\sum\sum_{i \neq j} \sigma_{ij}. \tag{1.17}$$

Dividing both sides of the resulting inequality by $\mathrm{var}\,(Y)$ finally gives the desired lower bound of the reliability of the composite measurement Y

$$R_m = \frac{\mathrm{var}\,(T)}{\mathrm{var}\,(Y)} = \frac{\sum_i \sum_j \sigma_{ij}}{\mathrm{var}\,(Y)} \geq \frac{m}{m-1} \frac{\sum \sum_{i \neq j} \mathrm{cov}(Y_i, Y_j)}{\mathrm{var}\,(Y)} = \alpha, \tag{1.18}$$

where the equality

$$\mathrm{cov}(Y_i, Y_j) = \mathrm{cov}(T_i, T_j) = \sigma_{ij} \qquad \text{for } i \neq j$$

is an easy consequence of the assumptions set at the beginning of Chapter 1.

To see when (1.14) holds as an equality, we need to know when the inequalities (1.15) and (1.16) hold as equalities. In (1.15), the equality holds for $\sigma_{ii} = \sigma_{jj}$. In (1.16), the equality holds if and only if $\mathrm{corr}(T_i, T_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = 1$. These two assumptions correspond with the assumptions for essential τ-equivalence, as shown in Lemma 1.4. □

To give a summary of Theorem 1.7: Cronbach's alpha is equal to the reliability of the composite measurement if and only if the measurement is composed of essentially τ-equivalent items. In other cases Cronbach's alpha is only a lower bound of the reliability!

Example 1.1. *For example, consider a measurement composed of m items whose true scores are not correlated at all:* $\mathrm{corr}(T_i, T_j) = 0$, *for $i \neq j$. In this case, the Cronbach's alpha is equal to zero, nevertheless the reliability*

$$R_m = \frac{\sum_j \mathrm{var}\,(T_j)}{\sum_j \mathrm{var}\,(T_j) + m\sigma_e^2}$$

can be close to 1 for σ_e^2 close to 0.

Example 1.2. *On the other hand, suppose a measurement composed of two items whose true scores are highly correlated:* $\mathrm{corr}(T_1, T_2) = 1$, *but* $\mathrm{var}\,(T_1) = 1$ *and* $\mathrm{var}\,(T_2) = 100$. *Then*

$$\alpha_{CR} \leq 2\frac{10 + 10}{10 + 10 + 1 + 100 + 2\sigma_e^2} \leq \frac{1}{3}$$

but the reliability R_2 of composite measurement $Y_1 + Y_2$

$$R_2 = \frac{10 + 10 + 1 + 100}{10 + 10 + 1 + 100 + 2\sigma_e^2}$$

can be close to 1 *for σ_e^2 close to zero.*

As shown above, Cronbach's alpha says more about items' *internal consistency* (how correlated the items' true scores are, whether they have the same variance). From this observation, it seems clear that the use of Cronbach's alpha for estimation of reliability should be done with caution.

1.4 Reliability in the ANOVA framework

Classical testing situations can be described in the framework of ANOVA (analysis of variance) models. In this section we discuss some of the models and we define the reliability of composite measurement for them. In more detail, the mixed effect model of a two-way ANOVA is studied where the reliability merges with Cronbach's alpha and with the intraclass correlation coefficient.

1.4.1 Mixed effect two-way ANOVA model

When studying properties of a certain educational test, we are usually interested in the *fixed set of items* the test is composed of. Also, we usually suppose that the *group of students* taking the test is a *random sample* of all possible test-takers. The testing situation can therefore be described by the **mixed effect model of a two-way ANOVA**: We assume that the score Y_{ij} reached by the i-th student in the j-th item can be expressed as

$$Y_{ij} = A_i + b_j + e_{ij} \qquad i = 1, \dots, n, \qquad j = 1, \dots, m, \tag{1.19}$$

where the ability of i-th person $A_i \sim \mathrm{N}(\mu, \sigma_A^2)$ is a random variable obeying the normal distribution, b_j is a fixed parameter describing j-th item's difficulty (we often require $\sum_j b_j = 0$), and $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$ is a normally distributed error term independent of abilities A_p for $p = 1, \dots, n$. Error terms e_{ij} are supposed to be mutually independent.

The reliability (defined as the ratio of the variance of the measured property and the variance of the observed score) of composite measurement $Y_i = \sum_j Y_{ij}$ is

$$R_m = \frac{\mathrm{var}\,(A_i)}{\mathrm{var}\,(Y_i)} = \frac{m^2\sigma_A^2}{m^2\sigma_A^2 + m\sigma_e^2} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{1}{m}\sigma_e^2}. \tag{1.20}$$

This is equal to Cronbach's alpha

$$\alpha_{CR} = \frac{m}{m-1}\frac{\sum\sum_{j\neq k}\mathrm{cov}(Y_{ij}, Y_{ik})}{\mathrm{var}\,(Y_i)} = \frac{m}{m-1}\frac{m(m-1)\sigma_A^2}{m^2\sigma_A^2 + m\sigma_e^2} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{1}{m}\sigma_e^2}.$$

Let us now look at the sums of squares and at their distributions:

$$SS_A = \sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2 \sim (m\sigma_A^2 + \sigma_e^2)\chi^2(n-1), \tag{1.21}$$

$$SS_e = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_{\bullet j} - \bar{Y}_{i\bullet} + \bar{Y}_{\bullet\bullet})^2 \sim \sigma_e^2\chi^2((n-1)(m-1)), \tag{1.22}$$

where the "bullet" notation was explained in (1.13). The mean squares MS_A and MS_e have mean values

$$\mathrm{E}\,MS_A = \mathrm{E}\,SS_A/(n-1) = m\sigma_A^2 + \sigma_e^2,$$
$$\mathrm{E}\,MS_e = \mathrm{E}\,SS_e/((n-1)(m-1)) = \sigma_e^2.$$

Therefore, the reliability (1.20) can be expressed as

$$R_m = \frac{\mathrm{E}MS_A - \mathrm{E}MS_e}{\mathrm{E}MS_A} = 1 - \frac{\mathrm{E}MS_e}{\mathrm{E}MS_A}. \tag{1.23}$$

As was already mentioned, this merges with Cronbach's alpha. Moreover, (1.23) is also the intraclass correlation coefficient (ICC) in this model. For further connections between ICC and Cronbach's alpha see Bravo and Potvin (1991).

When replacing mean values $\mathrm{E}MS_A, \mathrm{E}MS_e$ in (1.23) by their unbiased estimators MS_A, MS_e, we get an estimator for reliability

$$\hat{R}_m = 1 - \frac{MS_e}{MS_A}. \tag{1.24}$$

Since the mean squares MS_A and MS_e can be written as

$$\begin{aligned} MS_A &= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}, \\ MS_e &= \frac{1}{m-1}\sum_{j=1}^{m} s_{jj} - \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=1}^{m} s_{ij}, \end{aligned}$$

the estimate (1.24) merges with the sample estimate of Cronbach's alpha (1.11).

Further, (1.24) can be written as

$$\hat{R}_m = \hat{\alpha}_{CR} = 1 - \frac{1}{F_A}, \tag{1.25}$$

where F_A is a statistics used for testing hypothesis $H_0 : \sigma_A^2 = 0$ (either in mixed effect model (1.19) or in random effect model (1.26)). Statistics F_A has under hypothesis H_0 the Fisher-Snedecor distribution $F_{n-1,(n-1)(m-1)}$. We reject the hypothesis for large values of F_A. Therefore, (1.25) implies that the greater the estimate of reliability is, the better the educational test can distinguish between the students.

Also, (1.25) implies that $\hat{\alpha}_{CR} = 1$ in the case when $Y_{ij} = a_i + b_j$, where a_i and b_j are appropriate constants. In this case only one item would be enough to get all the information about a student!

Expression (1.25) should also warn us that the estimate $\hat{\alpha}_{CR}$ can take negative values, although only positive values make sense for reliability.

Finally, as a corollary of (1.25) we can derive the confidence interval for Cronbach's alpha and for reliability: (1.21) and (1.22) imply that the ratio

$$\frac{MS_A}{MS_e}\frac{\sigma_e^2}{(m\sigma_A^2 + \sigma_e^2)} = F_A \cdot (1 - R_m)$$

has Fisher-Snedecor distribution $F_{n-1,(n-1)(m-1)}$. Thus

$$\begin{aligned} \gamma/2 &= \mathrm{P}(F_A \cdot (1 - R_m) \geq F_{n-1,(n-1)(m-1)}(\gamma/2)) \\ &= \mathrm{P}\left(R_m \leq 1 - \frac{F_{n-1,(n-1)(m-1)}(\gamma/2)}{F_A}\right), \\ \gamma/2 &= \mathrm{P}(F_A \cdot (1 - R_m) \leq F_{n-1,(n-1)(m-1)}(1 - \gamma/2)) \\ &= \mathrm{P}\left(R_m \geq 1 - \frac{F_{n-1,(n-1)(m-1)}(1 - \gamma/2)}{F_A}\right), \end{aligned}$$

and the $(1-\gamma)100\%$ confidence interval for R_m is $\langle R_{min}, R_{max}\rangle$, where

$$\begin{aligned} R_{min} &= \max\left(0, 1 - \frac{F_{n-1,(n-1)(m-1)}\left(\frac{\gamma}{2}\right)}{F_A}\right), \\ R_{max} &= \min\left(1, 1 - \frac{F_{n-1,(n-1)(m-1)}\left(1-\frac{\gamma}{2}\right)}{F_A}\right). \end{aligned}$$

1.4.2 Other ANOVA models

When the test items are understood as a random sample from a bigger set of items, the **random effect two-way ANOVA model** should be used:

$$Y_{ij} = A_i + B_j + e_{ij} \qquad i = 1, \dots, n, \qquad j = 1, \dots, m. \tag{1.26}$$

The assumptions of this model are the same as in (1.19), only the j-th item's difficulty is supposed to be a normally distributed random variable $B_j \sim \mathrm{N}(0, \sigma_B^2)$, independent of A_i and e_{ip} for $i = 1, \dots, n$, $p = 1, \dots, m$.
The reliability of composite measurement $Y_i = \sum_j Y_{ij}$ in model (1.26) can be expressed as

$$R_m = \frac{\mathrm{var}\,(A_i)}{\mathrm{var}\,(Y_i)} = \frac{m^2\sigma_A^2}{m^2\sigma_A^2 + m\sigma_B^2 + m\sigma_e^2} = \frac{\sigma_A^2}{\sigma_A^2 + \frac{1}{m}\sigma_B^2 + \frac{1}{m}\sigma_e^2}. \tag{1.27}$$

This is not equal to Cronbach's alpha (1.10). Nevertheless, we can estimate the reliability (1.27) using the same principles as in the previous section. We can easily derive the mean values of the mean squares:

$$\begin{aligned} \mathrm{E}\,MS_A &= \mathrm{E}\,SS_A/(n-1) = m\sigma_A^2 + \sigma_e^2, \\ \mathrm{E}\,MS_B &= \mathrm{E}\,SS_B/(m-1) = n\sigma_B^2 + \sigma_e^2, \\ \mathrm{E}\,MS_e &= \mathrm{E}\,SS_e/((n-1)(m-1)) = \sigma_e^2. \end{aligned}$$

The reliability (1.27) can therefore be expressed as

$$R_m = \frac{\mathrm{E}\,MS_A - \mathrm{E}\,MS_e}{\mathrm{E}\,MS_A + \frac{\mathrm{E}\,MS_B - \mathrm{E}\,MS_e}{n}}. \tag{1.28}$$

The unbiased estimators of variances σ_A^2, σ_B^2 and σ_e^2 are

$$\begin{aligned} \hat{\sigma}_e^2 &= MS_e, \\ \hat{\sigma}_A^2 &= \frac{MS_A - MS_e}{m}, \\ \hat{\sigma}_B^2 &= \frac{MS_B - MS_e}{n}, \end{aligned}$$

and the reliability of the total score (1.28) can be estimated similarly to (1.25) by

$$\hat{R}_m = \frac{MS_A - MS_e}{MS_A + \frac{MS_B - MS_e}{n}}. \tag{1.29}$$

Higher order ANOVA models can be used when variation arises from more sources (e.g. variation due to administration, etc.). The sources of variation (the *effects*) might be *fixed* or *random*, as was shown with item effect in models (1.19) and (1.26). The ANOVA model might be *full* (as was the case in models (1.19) and (1.26) where all the students answered all of the items) or the model might be *nested* (e.g. the case of k administrations, where each student took the test only within one of the possible administrations).

When evaluating the reliability of measurement, we should first identify all the sources of the variation – the effects. Then we should identify whether these effects are fixed or random and whether the model is full or nested. Further, the *true value* of the measured property we are interested in should be identified. Then, the reliability can be expressed as a ratio of the true score variance to the total variance of the measurement. Finally, the reliability can be estimated using mean squares as shown within models (1.19) and (1.26): We express the reliability in terms of mean values of mean squares (see (1.23), (1.28)). The estimate is then formed by the mean squares themselves (see (1.24), (1.29)).

At the end of this section, let us remind ourselves of the **least squares estimators of parameters** in the ANOVA models. Let us for example suppose the two-way fixed effect model

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, m, \tag{1.30}$$

where μ, α_i and β_j are fixed, $\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{m} \beta_j = 0$ and $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$. The estimators are gained by minimizing the expression

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (Y_{ij} - (\mu + \alpha_i + \beta_j))^2$$

with respect to i and j. We get

$$\hat{\mu} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} y_{ij}}{mn} = \bar{y}_{\bullet\bullet}, \tag{1.31}$$

$$\hat{\alpha}_i = \frac{\sum_{j=1}^{m} y_{ij}}{m} - \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} y_{ij}}{mn} = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}, \tag{1.32}$$

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} y_{ij}}{n} - \frac{\sum_{j=1}^{m} \sum_{i=1}^{n} y_{ij}}{mn} = \bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}. \tag{1.33}$$

These estimators are intuitive ones: the i-th student ability α_i is estimated by his/her proportion of correct answers in the whole test and the j-th item difficulty β_j is estimated by the proportion of the correct responses to that item from the responses of all the students (standardized, so that the mean ability and mean difficulty are zero). We denote these estimators as **conventional** and we will return to them in Section 2.1.2.

Chapter 2

Measurements consisting of dichotomously scored items

The theory discussed in Chapter 1 is hardly applicable if the test is composed of dichotomously scored items. Model (1.1), supposing that the measurement Y is a sum of the true value T and an error term e, is misleading in the case when Y reaches only 0 or 1.

In this chapter we describe some models which could be appropriate for the case of dichotomously scored items. As in Chapter 1, we pay the most attention to the mixed effect models, assuming that the items are fixed and that the students are a random sample from the population. We generally assume that the score Y_{ij} reached by the i-th student on the j-th item is either 0 (wrong answer) or 1 (correct answer), and that it has a distribution

$$Y_{ij} \sim f_{ij}(\bullet; A_i, b_j) \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, m, \tag{2.1}$$

where A_i's (describing students' ability) are independently distributed with the same distribution function $H(A_i)$. Moreover, we assume that Y_{ij} and $Y_{ij'}$ are conditionally independent given A_i.

In Section 2.1 we discuss the model used most often for description of dichotomous items – the Rasch model. Attention is paid mainly to methods of parameter estimation.

In a certain stage of our research, we were not able to define the reliability of measurement composed of items obeying the Rasch model. On the other hand, the reliability is easily defined by the intraclass correlation coefficient (ICC) for binary data obeying common correlation models, such as the beta-binomial model. Nevertheless, the beta-binomial model is not appropriate for our situation since it does not allow for different item difficulties. Therefore, as a step aside, we tried to modify the beta-binomial model. The resulting *modified beta-binomial model* is discussed in Section 2.2. Finally, Section 2.3 brings an extension of this model to other ability distributions.

2.1 Rasch model

The model used most often for describing dichotomously scored items (often in the context of the Item Response Theory) is the logit-normal model, called the Rasch model (see Rasch (1960)). In the Rasch model, the probability of correct response $y_{ij} = 1$ or false response $y_{ij} = 0$ of person i on item j is given by[1]:

$$\mathrm{P}(Y_{ij} = y_{ij}|A_i) = \frac{\exp[y_{ij}(A_i + b_j)]}{1 + \exp(A_i + b_j)}, \tag{2.2}$$

where $A_i \sim \mathrm{N}(\mu, \sigma_A^2)$ describes the level of ability of person i, and b_j is an unknown parameter describing the difficulty of item j. The conditional distributions are assumed to be independent. Usually, we require $\sum_j b_j = 0$.

A direct generalization of the Rasch model is the three parameter logistic model, where item properties are described by three parameters: besides *difficulty parameter* b_j, there is a *discrimination parameter* c_j describing discrimination power of the item and a *guessing parameter* d_j describing the probability of guessing that item by a person with no knowledge. The probability of a correct response of person i on item j is given by

$$P(Y_{ij} = y_{ij}; A_i, b_j, c_j, d_j) = d_j + (1 - d_j)\frac{\exp[y_{ij}c_j(A_i + b_j)]}{1 + \exp[c_j(A_i + b_j)]}. \tag{2.3}$$

In this model, a nice and clear interpretation of parameters is possible: Let us define the item characteristic curve of item j as $f_j(a) = P[Y_{ij} = 1|a, b_j, c_j, d_j]$. After estimating the parameters of an item, the item characteristic curve can be plotted out (see Figure 2.1). By further analysis of the function $f_j(a)$ we can derive that:

- If $c_j > 0$ then $f_j(a)$ is increasing (so that the better students are more likely to answer the item correctly), which is a reasonable assumption for an item.
- If $c_j > 0$, then $d_j = \lim_{a\to-\infty} f_j(a)$, therefore d_j describes the probability that person without any knowledge answers the item correctly.
- Difficulty parameter b_j can be understood as a specific value on the ability scale: If a person has ability $a = -b_j$, then the probability that the person answers item j correctly is $\frac{1+d_j}{2}$, and so this probability is exactly in the middle between 1 and d_j.
- The first derivative of function f_j at point $-b_j$ is equal to $c_j\frac{1-d_j}{4}$, thus the discrimination parameter c_j is described by the slope of item characteristic curve at point $-b_j$, more precisely it is equal to $f'(-b_j)\frac{4}{1-d_j}$.

The item characteristic curve describes the properties of an item very clearly: we can easily read its difficulty, the probability that persons with no knowledge

[1] In publications, often a slightly reparametrized model is called the Rasch model: there is a minus sign in front of the item effect b_j in (2.2). In such a case, b_j can be understood as "item difficulty", while in (2.2) it is rather "item easiness". We use notation (2.2) for better comparison with model (1.19).

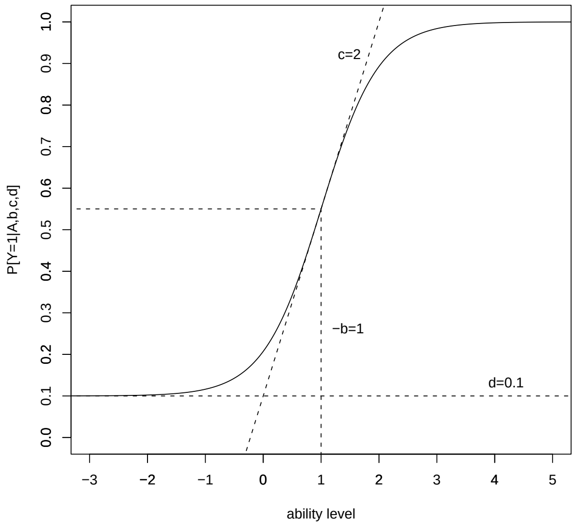

Figure 2.1: Item characteristic curve for an item with difficulty parameter $b_j = -1$, discrimination parameter $c_j = 2$ and guessing parameter $d_j = 0.1$

answer it correctly. From the item characteristic curve plotted in Figure 2.1 we can easily see, for example, that the described item can very well distinguish between the students with ability level between 0 and 2. On the other hand, this item does not distinguish very well between students with lower ability level, nor between students with higher ability level.

Other extensions of the Rasch model are possible, too. Among these there are the extensions to polytomous models, such as the partial credit model, rating scale model, binomial trials and Poisson counts model. The majority of these models can be covered in a generalized linear model. Other possible extensions are models for items in which response time or number of successful attempts are recorded. A well-arranged overview of extensions of the Rasch model can be found in van der Linden and Hambleton (1997). An advantage of models containing more parameters is a better description of the situation. A disadvantage is that with small sample sizes it may result in unstable estimators of item parameters.

2.1.1 Parameter estimation – ML methods

Let us for the rest of this section concentrate on the Rasch model (2.2). There are three likelihood–based methods available for an item parameter estimation: joint maximum likelihood (JML), marginal maximum likelihood (MML) and conditional maximum likelihood (CML). All three of the algorithms based on maximum likelihood described in the next three subsections use the iterative procedures. Therefore, the estimation procedures for item parameters can be hard to explain to non-statisticians. Nevertheless, in Section 2.1.2 we show that the conventional estimator (1.33) can be understood as an approximation of the estimators resulting from ML methods studied in the following text.

Joint maximum likelihood

The joint likelihood function for one-parameter Rasch model (2.2) is given by

$$p(\boldsymbol{y};\boldsymbol{\omega}) = \prod_{i=1}^{n}\prod_{j=1}^{m} P(Y_{ij} = y_{ij}; a_i, b_j), \tag{2.4}$$

with $\boldsymbol{\omega} = (\boldsymbol{b}^T, \boldsymbol{a}^T)$, $\boldsymbol{a} = (a_1, \dots, a_n)$ being the vector of abilities, $\boldsymbol{b} = (b_1, \dots, b_m)$ representing the vector of difficulties of items and with $P(Y_{ij} = y_{ij}; a_i, b_j)$ given by (2.2). We should emphasize that, when studying the joint likelihood function, we in fact work with a model in which the impact of the i-th student a_i is considered to be a **fixed effect.** The item parameters are estimated by maximizing (2.4) with respect to $\boldsymbol{\omega}$ given the data $\boldsymbol{x}$. Nevertheless, when keeping the number of item parameters and increasing the number of tested persons, the method leads to inconsistent estimators. This is caused by the fact that a limited number of parameters of interest (item difficulties $\boldsymbol{b}$) is to be estimated in the presence of the growing number of nuisance parameters (abilities $\boldsymbol{a}$). A wide discussion of this problem in a general setting was done already in Neyman and Scott (1948). Eliminating the nuisance parameters gives a solution to this problem. The elimination can be accomplished by the marginal or the conditional maximum likelihood method.

Marginal maximum likelihood

When estimating the item parameters using the marginal maximum likelihood (MML) method, we usually assume that the abilities $\boldsymbol{A}$ constitute a random sample from an ability distribution with density $h(A;\boldsymbol{\xi})$, with $\boldsymbol{\xi}$ the parameters of the ability distribution. The joint probability can be then written as

$$p(\boldsymbol{y};\boldsymbol{b},\boldsymbol{\xi}) = \prod_{i=1}^{n}\int_{-\infty}^{\infty}\prod_{j=1}^{m} P(Y_{ij} = y_{ij}|A_i; b_j)h(A_i;\boldsymbol{\xi})dA_i, \tag{2.5}$$

with $P(Y_{ij} = y_{ij}|A_i; b_j)$ again given by (2.2). The above mentioned marginal likelihood function is maximized with respect to $\boldsymbol{b}$ and $\boldsymbol{\xi}$. The nuisance parameters are eliminated by integrating over them. Often, the ability distribution is considered to be normal with unknown parameters μ_A and σ_A^2, which are estimated together with $\boldsymbol{b}$. The main problem of this method is the correct specification of the ability distribution. If the distribution is not specified correctly, the method can lead to biased estimators of item parameters. The MML method can be used also without specifying a parametric form of the ability distribution. This nonparametric distribution is then estimated together with the item parameters. EM algorithm and MCMC method can be used for the estimation.

Conditional maximum likelihood

The last approach to item parameter estimation is the conditional maximum likelihood (CML) method. It results from the fact that if there exist sufficient statistic for the nuisance parameters, the model can be separated in a conditional part dependent only on the parameters of interest and a part which models the sufficient statistic.

Lemma 2.1. *Let us define the total score variables*[2]

$$S_i = \sum_{j=1}^{m} Y_{ij} \qquad i = 1, \dots, n,$$

and their realizations

$$s_i = \sum_{j=1}^{m} y_{ij} \qquad i = 1, \dots, n.$$

Then under the Rasch model (2.2), S_i is a sufficient statistic for a_i, $i = 1, \dots, n$.

Proof: Let us denote $(\boldsymbol{x}|s_i)$ any vector $\boldsymbol{x} = (x_1, \dots, x_m)$ for which it holds that $\sum_{j=1}^{m} x_j = s_i$. Let us denote $\psi_i = \exp a_i$, and $\epsilon_j = \exp b_j$. Then, supposing the model (2.2), the conditional probability $P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{S} = \boldsymbol{s})$ can be rewritten as

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{S} = \boldsymbol{s}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{S} = \boldsymbol{s})}{P(\boldsymbol{S} = \boldsymbol{s})} = \frac{\prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\psi_i^{y_{ij}} \epsilon_j^{y_{ij}}}{1+\psi_i \epsilon_j}}{\prod_{i=1}^{n} \sum_{(\boldsymbol{x}|s_i)} \prod_{j=1}^{m} \frac{\psi_i^{x_{ij}} \epsilon_j^{x_{ij}}}{1+\psi_i \epsilon_j}}$$

$$= \prod_{i=1}^{n} \frac{\prod_{j=1}^{m} \frac{1}{1+\psi_i\epsilon_j} \psi_i^{s_i} \prod_{j=1}^{m} \epsilon_j^{y_{ij}}}{\prod_{j=1}^{m} \frac{1}{1+\psi_i\epsilon_j} \psi_i^{s_i} \sum_{(\boldsymbol{x}|s_i)} \prod_{j=1}^{m} \epsilon_j^{x_{ij}}} = \prod_{i=1}^{n} \frac{\prod_{j=1}^{m} \epsilon_j^{y_{ij}}}{\sum_{(\boldsymbol{x}|s_i)} \prod_{j=1}^{m} \epsilon_j^{x_{ij}}},$$

for $\boldsymbol{S} = \boldsymbol{s}$ and as 0 otherwise. Therefore the conditional probability $P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{S} = \boldsymbol{s})$ does not depend on a_i, $i = 1, \dots, n$, and thus the total score S_i is a sufficient statistic for a_i, $i = 1, \dots, n$. □

Since in the Rasch model (2.2) the total score S_i is a sufficient statistic for a_i, $i = 1, \dots, n$, the likelihood function (2.4) can be rewritten as:

$$p(\boldsymbol{y}; \boldsymbol{\omega}) = \prod_{i=1}^{n} f(\boldsymbol{y_i}|s_i; \boldsymbol{b}) \prod_{i=1}^{n} g(s_i; \boldsymbol{b}; a_i), \tag{2.6}$$

with $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{im})$ the response vector of person i. Maximization of the conditional likelihood

$$\prod_{i=1}^{n} f(\boldsymbol{y_i}|s_i; \boldsymbol{b}) \tag{2.7}$$

with respect to $\boldsymbol{b}$ leads under mild conditions to consistent and asymptotically normally distributed estimates (see Andersen (1970)).

[2] While in the rest of the work we denote the total score of the i-th student Y_i, for better clarity of this section we choose a different symbol here: S_i.

An interesting topic in CML estimates is their efficiency. The problem is that when estimating the item parameters, only the conditional likelihood (2.7) is used and the second part of the full likelihood (2.6), the marginal distribution of S, is neglected. Nevertheless, this second part could possibly contain some information on the item parameters. For evaluating the loss of information due to using the CML method, the F-information can be defined. This is a generalization of Fisher information matrix for the case when a part of the parameters is nuisance. The properties of F-information and the loss of information in CML estimation is in detail studied in Eggen (2000).

2.1.2 Conventional estimator – approximation of MLE

When one is asked to estimate difficulty of an item, probably the simplest thing he/she can think of is the proportion of correct responses to that item. In this subsection we would like to show that this conventional estimator is justified and that it approximates the estimators mentioned above.

We will demonstrate our claim on the joint maximum likelihood method, therefore the impact of the student ability a_i is again considered to be a **fixed effect.** Let us make a slight reparametrization of the Rasch model:

$$P[Y=1|a_i,b_j] = \frac{e^{a_i+b_j}}{1+e^{a_i+b_j}} = \frac{e^{\mu+\alpha_i+\beta_j}}{1+e^{\mu+\alpha_i+\beta_j}} \overset{def}{=} f(\mu+\alpha_i+\beta_j), \tag{2.8}$$

with $\sum \alpha_i = \sum \beta_j = 0$. Let us consider the Taylor approximation

$$f(\mu+\alpha_i+\beta_j) \doteq f(\mu) + f'(\mu)(\alpha_i+\beta_j) = f(\mu) + f(\mu)(1-f(\mu))(\alpha_i+\beta_j). \tag{2.9}$$

Let us define $\eta = f(\mu) = \frac{e^\mu}{1+e^\mu}$ and

$$f(\mu+\alpha_i+\beta_j) \doteq \eta + \eta(1-\eta)(\alpha_i+\beta_j) \overset{def}{=} K_{ij}, \tag{2.10}$$

then the new joint likelihood function can be written as

$$L = \prod_{i=1}^{n}\prod_{j=1}^{m} K_{ij}^{y_{ij}}(1-K_{ij})^{(1-y_{ij})}, \tag{2.11}$$

and its logarithm can be written as

$$\ln L = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\ln K_{ij} + (1-y_{ij})\ln(1-K_{ij}). \tag{2.12}$$

Let us now maximize the logarithm of the joint likelihood function with respect to η, α_i and β_j to get the maximum likelihood estimators $\hat{\eta}$, $\hat{\alpha_i}$, $\hat{\beta_j}$:

$$0 = \frac{\partial \ln L}{\partial \eta} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[y_{ij}\frac{1+(1-2\eta)(\alpha_i+\beta_j)}{\eta+\eta(1-\eta)(\alpha_i+\beta_j)} + (1-y_{ij})\frac{-1-(1-2\eta)(\alpha_i+\beta_j)}{1-\eta-\eta(1-\eta)(\alpha_i+\beta_j)}\right]$$

$$0 = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\frac{y_{ij}}{\eta+\eta(1-\eta)(\alpha_i+\beta_j)} - \frac{(1-y_{ij})}{1-\eta-\eta(1-\eta)(\alpha_i+\beta_j)}\right]$$

$$0 = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}(1-\eta-\eta(1-\eta)(\alpha_i+\beta_j)) - (1-y_{ij})(\eta+\eta(1-\eta)(\alpha_i+\beta_j)).$$

When using reparametrization conditions $\sum \alpha_i = \sum \beta_j = 0$, we get the maximum likelihood estimator

$$\hat{\eta} = \bar{y}_{\bullet\bullet}, \tag{2.13}$$

where the "bullet" notation was explained in (1.13). Similarly for ability parameters α_i we get

$$\begin{aligned}
0 &= \frac{\partial \ln L}{\partial \alpha_i} = \sum_{j=1}^{m} \left[\frac{y_{ij}\eta(1-\eta)}{\eta + \eta(1-\eta)(\alpha_i+\beta_j)} - \frac{(1-y_{ij})\eta(1-\eta)}{1-\eta-\eta(1-\eta)(\alpha_i+\beta_j)} \right] \\
0 &= \sum_{j=1}^{m} y_{ij}(1-\eta-\eta(1-\eta)(\alpha_i+\beta_j)) - (1-y_{ij})(\eta+\eta(1-\eta)(\alpha_i+\beta_j)), \\
0 &= y_{i\bullet} - m\eta - m\eta(1-\eta)\alpha_i,
\end{aligned}$$

therefore we have

$$\hat{\alpha_i} = \frac{\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}}{\bar{y}_{\bullet\bullet}(1-\bar{y}_{\bullet\bullet})}. \tag{2.14}$$

Similarly for item-difficulty parameters

$$\begin{aligned}
0 &= \frac{\partial \ln L}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{y_{ij}\eta(1-\eta)}{\eta + \eta(1-\eta)(\alpha_i+\beta_j)} - \frac{(1-y_{ij})\eta(1-\eta)}{1-\eta-\eta(1-\eta)(\alpha_i+\beta_j)} \right], \\
0 &= \sum_{i=1}^{n} y_{ij}(1-\eta-\eta(1-\eta)(\alpha_i+\beta_j)) - (1-y_{ij})(\eta+\eta(1-\eta)(\alpha_i+\beta_j)), \\
0 &= y_{\bullet j} - n\eta + n\eta(1-\eta)\beta_j.
\end{aligned}$$

Therefore, the maximum likelihood estimator for item difficulty is

$$\hat{\beta_j} = \frac{\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet}}{\bar{y}_{\bullet\bullet}(1-\bar{y}_{\bullet\bullet})}. \tag{2.15}$$

Estimators (2.14) and (2.15) are multiples of the conventional estimators (1.32) and (1.33) derived for model of analysis of variance in Section 1.4. In this sense, the conventional estimators can be understood as justified approximations of the estimators based on the Rasch model.

2.2 Beta-binomial model and its modification

An often used model in reliability studies of binary data (see for example Ridout et al. (1999), Zou and Donner (2004)) is the **beta-binomial model**. In this model, we assume that the probability of success π_i varies over subjects $i = 1, \dots, n$ according to a beta distribution with parameters a and b, and conditional to this probability, the total score Y_i of the i-th person is binomially distributed. The choice of beta distribution for π_i is logical since it is a flexible distribution and leads to mathematically tractable results. The beta probability density function is

$$f(\pi; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1}(1-\pi)^{b-1}, \qquad 0 \leq \pi \leq 1,$$

where $a > 0$ and $b > 0$. The beta distribution has mean and variance equal to

$$\mathrm{E}\,(\pi) = \mu, \qquad \mathrm{var}\,(\pi) = \mu(1-\mu)\theta/(1+\theta) = \mu(1-\mu)\rho,$$

where $\mu = \frac{a}{a+b}$ is the marginal probability of success for any individual, $\theta = \frac{1}{a+b}$ and ρ is the intraclass correlation coefficient $\rho = \mathrm{corr}(Y_{ij}, Y_{ik})$, $j \neq k$ common for any subject and any pair of responses (this model is therefore sometimes called the common correlation model).

Marginally, averaging with respect to the beta distribution for π_i, the total score Y_i has the beta-binomial distribution with

$$\mathrm{P}(Y_i = y_i) = \binom{n}{y_i} \frac{\prod_{j=0}^{y_i-1}(\mu + j\theta) \prod_{j=0}^{n-y_i-1}(1-\mu+j\theta)}{\prod_{j=0}^{n-1}(1+j\theta)}, \tag{2.16}$$

and with the first two moments equal to

$$\mathrm{E}\,(Y_i) = n\mu = n\frac{a}{a+b},$$

$$\mathrm{var}\,(Y_i) = n\mu(1-\mu)\left[1 + (n-1)\frac{\theta}{1+\theta}\right].$$

An unpleasant property of this model for our situation is the fact that it does not allow for different difficulties of items. Hand in hand with this goes the common correlation structure.

When trying to extend the beta-binomial model to cover different difficulties of items and yet to preserve the model structure, we can think of the following **modified beta-binomial model:** We again assume that the probability of success π_i varies over subjects $i = 1, \ldots, n$ according to a beta distribution with parameters a and b. We quantify the impact of the difficulty of the j-th item by a small number b_j, assuming that $\sum_{j=1}^m b_j = 0$. When parameters a, b of the beta distribution are large enough, there is a slight danger that the sums $\pi_i + b_j$ fall out of the interval $(0,1)$. Therefore, when neglecting the probability that the sums $\pi_i + b_j$ fall out of the interval $(0,1)$, $Y_{i1}, \ldots, Y_{im}$ are for a given π_i independent random variables with Bernoulli distribution with probability of success $(\pi_i + b_j)$. To be more exact, let us suppose

$$\begin{aligned} \pi_i \sim B(a,b) \qquad & a > 0, \quad b > 0, \\ b_j \text{ fixed, such that} \qquad & \sum_{j=1}^m b_j = 0, \\ \mathrm{P}(Y_{ij} = 1|\pi_i) = \max(\min(\pi_i + b_j, 1), 0) \qquad & i = 1, \ldots, n \quad j = 1, \ldots, m. \end{aligned} \tag{2.17}$$

Then we say that the total scores $Y_i = \sum_{j=1}^m Y_{ij}$ obey the modified beta-binomial model. Let us now look at the first two conditional and unconditional moments of Y_{ij} in the modified beta-binomial model.

Lemma 2.2. *Let us suppose that* Y_i, $i = 1, \dots n$ *obey the modified beta-binomial model (2.17) and let us neglect the small probability that* $\pi_i + b_j$ *fall out of the interval* $(0, 1)$. *Then for conditional mean and variance, it holds that*

$$\begin{aligned} \mathrm{E}\,(Y_{ij}|\pi_i) &= \pi_i + b_j, \\ \mathrm{var}\,(Y_{ij}|\pi_i) &= (\pi_i + b_j)(1 - (\pi_i + b_j)), \end{aligned}$$

and for unconditional mean and variance, it holds that

$$\begin{aligned} \mathrm{E}\,(Y_{ij}) &= \mu + b_j, \\ \mathrm{var}\,(Y_{ij}) &= \mu(1 - \mu) + b_j(1 - 2\mu - b_j). \end{aligned}$$

Finally, the covariances and correlations between Y_{ij} *and* $Y_{it,}$ *for* $j \neq t$ *are equal to*

$$\begin{aligned} \mathrm{cov}(Y_{ij}, Y_{it}) &= \mathrm{var}\,(\pi_i) = \rho\mu(1 - \mu), \\ \mathrm{corr}(Y_{ij}, Y_{it}) &= \rho \frac{1}{\sqrt{C_j C_t}}, \end{aligned}$$

where

$$\rho = \frac{\mathrm{var}\,\pi_i}{\mu(1 - \mu)} = \frac{1}{1 + a + b}$$

is the correlation between Y_{ij} *and* Y_{it}, $j \neq t$, *in the beta-binomial model, and*

$$C_j = 1 + b_j \frac{1 - 2\mu - b_j}{\mu(1 - \mu)}.$$

Proof: For conditional mean and variance, it holds that

$$\begin{aligned} \mathrm{E}\,(Y_{ij}|\pi_i) &= \mathrm{E}\,(Y_{ij}^2|\pi_i) = \mathrm{P}(Y_{ij} = 1|\pi_i) = \pi_i + b_j, \\ \mathrm{var}\,(Y_{ij}|\pi_i) &= \mathrm{E}\,(Y_{ij}^2|\pi_i) - (\mathrm{E}\,(Y_{ij}|\pi_i))^2 = (\pi_i + b_j)(1 - (\pi_i + b_j)). \end{aligned}$$

Therefore the unconditional mean is

$$\mathrm{E}\,(Y_{ij}) = \mathrm{E}\,\mathrm{E}\,(Y_{ij}|\pi_i) = \mu + b_j,$$

where we assigned $\mu = a/(a + b)$ for the mean value of the beta distribution. For the unconditional variance, it holds that

$$\begin{aligned} \mathrm{var}\,(Y_{ij}) &= \mathrm{var}\,(\mathrm{E}\,(Y_{ij}|\pi_i)) + \mathrm{E}\,(\mathrm{var}\,(Y_{ij}|\pi_i)) \\ &= \mathrm{var}\,(\pi_i + b_j) + \mathrm{E}\,((\pi_i + b_j)(1 - (\pi_i + b_j))) \\ &= \mathrm{E}\,\pi_i^2 - (\mathrm{E}\,\pi_i)^2 + \mathrm{E}\,(\pi_i + b_j - \pi_i^2 - \pi_i b_j - \pi_i b_j - b_j^2) \\ &= -\mu^2 + \mu + b_j - 2b_j\mu - b_j^2 \\ &= \mu(1 - \mu) + b_j(1 - 2\mu - b_j). \end{aligned}$$

The covariance of variables Y_{ij}, Y_{it} for $j \neq t$ equals

$$\begin{aligned} \mathrm{cov}(Y_{ij}, Y_{it}) &= \mathrm{cov}(\mathrm{E}\,(Y_{ij}|\pi_i), \mathrm{E}\,(Y_{it}|\pi_i)) \\ &= \mathrm{cov}(\pi_i + b_j, \pi_i + b_t) = \mathrm{var}\,(\pi_i) \end{aligned}$$

Finally, the correlation between Y_{ij} and Y_{it} for $j \neq t$ is

$$\begin{aligned}
\operatorname{corr}(Y_{ij}, Y_{it}) &= \frac{\operatorname{cov}(Y_{ij}, Y_{it})}{\sqrt{\operatorname{var} Y_{ij} \operatorname{var} Y_{it}}} \\
&= \frac{\operatorname{var} \pi_i}{\sqrt{\mu(1-\mu) + b_j(1-2\mu-b_j)}\sqrt{\mu(1-\mu) + b_t(1-2\mu-b_t)}} \\
&= \frac{\operatorname{var} \pi_i}{\mu(1-\mu)\sqrt{1+\frac{b_j(1-2\mu-b_j)}{\mu(1-\mu)}}\sqrt{1+\frac{b_t(1-2\mu-b_t)}{\mu(1-\mu)}}} \\
&= \rho \frac{1}{\sqrt{C_j C_t}}.
\end{aligned}$$

□

For constant difficulties of items $b_j = 0$ we get the common correlation structure, $\operatorname{corr}(Y_{ij}, Y_{it}) = \rho$. For unequal difficulties of items it is natural to assume $a = b$ (to assume symmetric distribution of knowledge), therefore $\mu = 1/2$. In this case $C_j = 1 - 4b_j^2$, thus the impact of $b_j < 1$ is small. Let us now study the first two moments of the total score $Y_i = \sum_{j=1}^m Y_{ij}$.

Lemma 2.3. *Let us suppose that Y_i, $i = 1, \dots n$ obey model (2.17) and let us neglect the small probability that $\pi_i + b_j$ fall out of the interval $(0, 1)$. Then for the first two conditional moments of the total score $Y_i = \sum_{j=1}^m Y_{ij}$, it holds that*

$$\begin{aligned}
\operatorname{E}(Y_i|\pi_i) &= m\pi_i, \\
\operatorname{var}(Y_i|\pi_i) &= m\pi_i(1-\pi_i) - m\kappa_b,
\end{aligned}$$

where $\kappa_b = \frac{1}{m}\sum_{j=1}^m b_j^2$, and the first two unconditional moments of the total score can be expressed as

$$\begin{aligned}
\operatorname{E}(Y_i) &= m\mu, \\
\operatorname{var}(Y_i) &= m\mu(1-\mu)(1+(m-1)\rho) - m\kappa_b.
\end{aligned}$$

Proof: In the modified beta-binomial model, we have

$$\begin{aligned}
\operatorname{E}(Y_i|\pi_i) &= m\frac{1}{m}\sum_{j=1}^m (\pi_i + b_j) = m\pi_i, \\
\operatorname{var}(Y_i|\pi_i) &= \operatorname{var}\left(\sum_{j=1}^m Y_{ij}|\pi_i\right) = \sum_{j=1}^m (\pi_i + b_j)(1 - (\pi_i + b_j)) \\
&= \sum_{j=1}^m \left(\pi_i + b_j - (\pi_i + b_j)^2\right) = m\pi_i - m\pi_i^2 - \sum_{j=1}^m b_j^2 \\
&= m\pi_i(1-\pi_i) - m\kappa_b,
\end{aligned}$$

where $\kappa_b = \frac{1}{m}\sum_{j=1}^m b_j^2$. Therefore, it holds that

$$\begin{aligned}
\mathrm{E}\,(Y_i) &= \mathrm{E}\,\mathrm{E}\,(Y_i|\pi_i) = m\mathrm{E}\,(\pi_i) = m\frac{a}{a+b} = m\mu, \\
\mathrm{var}\,(Y_i) &= \mathrm{var}\,(m\pi_i) + \mathrm{E}\,(m\pi_i(1-\pi_i) - m\kappa_b) \\
&= m^2\mathrm{var}\,(\pi_i) + m\mathrm{E}\,(\pi_i) - m\mathrm{E}\,(\pi_i^2) - m\kappa_b + m\mathrm{E}\,\pi_i^2 - m\mathrm{E}\,\pi_i^2 \\
&= (m^2 - m)\mathrm{var}\,\pi_i + m\mu + m\mu - m\mu^2 - m\kappa_b \\
&= (m^2 - m)\rho\mu(1-\mu) + m\mu(1-\mu) - m\kappa_b \\
&= m\mu(1-\mu)(1 + (m-1)\rho) - m\kappa_b.
\end{aligned}$$

□

The formulas derived in Lemmas 2.2 and 2.3 will be needed for deriving the formula of reliability in the modified beta binomial model (see Chapter 4).

2.3 General model with additive item effect

When going through the proofs of Lemma 2.2 and of Lemma 2.3, we can notice that the same formulas for conditional and unconditional means and variances would be applicable even if the distribution of π_i was not the beta-binomial distribution, but any other distribution on $(0, 1)$. To be more specific, let

$$\begin{aligned}
\mathrm{P}(\pi_i \le x) = F(x), &\quad \text{F being a distribution function on } (0,1), \\
\mathrm{E}\,\pi_i = \mu \in (0,1), &\quad \mathrm{var}\,\pi_i = \sigma_\pi^2, \\
b_j \text{ fixed, such that} &\quad \sum_{j=1}^m b_j = 0, \\
\mathrm{P}(Y_{ij} = 1|\pi_i) = \max(\min(\pi_i + b_j, 1), 0) &\quad i = 1,\dots,n \quad j = 1,\dots,m. \qquad (2.18)
\end{aligned}$$

Then we say that the total score $Y_i = \sum_{j=1}^m Y_{ij}$ obeys the general model with additive item effect.

Lemma 2.4. *The formulas for conditional and unconditional means and variances derived in Lemmas 2.2 and 2.3 also hold for the general additive model (2.18). In particular (formulas needed in Chapter 4): When neglecting the small probability that $\pi_i + b_j$ fall out of the interval $(0, 1)$, then it holds that*

$$\begin{aligned}
\mathrm{E}\,(Y_{ij}|\pi_i) &= \pi_i + b_j, \\
\mathrm{var}\,(Y_{ij}) &= \mu(1-\mu) + b_j(1 - 2\mu - b_j),
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\,(Y_i|\pi_i) &= m\pi_i, \\
\mathrm{var}\,(Y_i) &= m\mu(1-\mu)(1 + (m-1)\rho) - m\kappa_b,
\end{aligned}$$

where $\kappa_b = \frac{1}{m}\sum_{j=1}^m b_j^2$ and $\rho = \sigma_\pi^2/\mu(1-\mu)$ is the correlation between Y_{ij} and Y_{it}, $j \neq t$, in the model (2.18), where $b_j = 0$ for all $j = 1, \dots, m$.

Proof: The proof is identical with the proof of Lemmas 2.2 and 2.3. □

Example (Modified Rasch model)
As in the model (2.2), we suppose that the ability of the i-th student is $A_i \sim \mathrm{N}(\mu_A, \sigma_A^2)$ and that the difficulty of the j-th item can be described by parameters b_j. In the modified Rasch model, we suppose that the overall probability that the i-th student answers correctly to an item is

$$\pi_i = \operatorname{logit} A_i = \frac{\exp A_i}{1 + \exp A_i}$$

and that the effect of an item is additive, that is

$$\mathrm{P}(Y_{ij} = 1 | A_i) = \max(\min(\pi_i + b_j, 1), 0) \qquad i = 1, \dots, n \quad j = 1, \dots, m. \tag{2.19}$$

In other words, when neglecting the small probability that $\pi_i + b_j$ falls out of the interval $(0, 1)$, the probability that the i-th student answers correctly the j-th question is

$$\mathrm{P}(Y_{ij} = 1 | A_i) = \frac{\exp A_i}{1 + \exp A_i} + b_j.$$

This model can be understood as an approximation of the Rasch model (2.2).

Chapter 3

Common intraclass correlation

As follows from Lemma 1.1, even in the classical situation the reliability is closely associated with the correlation between two measurements of the same property. In papers devoted to binary data (see for example Ridout et al. (1999), Zou and Donner (2004)), mostly the reliability is understood to merge with the intraclass correlation. This is true for the class of models called *common correlation models*, where the correlation is common for all pairs of intraclass measurements: $\mathrm{corr}(Y_{ij}, Y_{ij'}) = \rho$, for all i and for all $j \neq j'$.

In Commenges and Jacqmin (1994), some of possibly equivalent definitions of the common correlation model are discussed. Since the general definition of reliability proposed in Chapter 4 of this thesis is inspired by the coefficient τ, proposed in paper of Commenges and Jacqmin, we devote this chapter to revision of the mentioned paper.

Commenges and Jacqmin study the random effect model

$$Y_{ij} \sim f_{ij}(\bullet; A_i) \qquad i = 1, \ldots, n, j = 1, \ldots, m \tag{3.1}$$

with finite variance, $\mathrm{var}\, Y_{ij} < \infty$, independently distributed effects A_i with the same distribution function $H(A_i)$ and with $Y_{ij}, Y_{ij'}$ conditionally independent given A_i. Our model (2.1) is a special case of (3.1): we assume $Y_{ij} = 0$ or 1 and we suppose that the model contains a fixed item effect b_j, common for all subjects $i = 1, \ldots, n$.

In the mentioned paper, two possible ways of defining the intraclass correlation coefficient are discussed. One is based on the correlation between two variables of the same group, $\mathrm{corr}(Y_{ij}, Y_{ij'})$, the other is based on a decomposition of the variance:

For $j \neq j'$, let $\rho_{ijj'} = \mathrm{corr}(Y_{ij}, Y_{ij'})$. If $\rho_{ijj'}$ does not depend on j and j', we define the *intraclass correlation coefficient for class* i by ρ_i. In more restricted models, all the ρ_i are equal to a *common intraclass coefficient* ρ.

Consider now the model (3.1). The variance of Y_{ij} can be decomposed as a function of its conditional moments:

$$\mathrm{var}\,(Y_{ij}) = \mathrm{E}\,[\mathrm{var}\,(Y_{ij}|A_i)] + \mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)]\,.$$

The first term of this decomposition is essentially the intragroup variance, that is, the part of the variance that is not due to variability of A_i. The second term is an intergroup variance, the part of the variance $\mathrm{var}\,(Y_{ij})$ that is due to variability of A_i. Commenges and Jacqmin further define

$$\tau_{ij} = \frac{\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)]}{\mathrm{var}\,(Y_{ij})}. \tag{3.2}$$

This coefficient is the relative part of the variance due to the effect of the variability of A_i on the conditional expectation of Y_{ij}, i.e., the heterogeneity of means between groups. Commenges and Jacqmin state that it may be the basis of a definition of an intraclass correlation. We add that, moreover, it may be the basis of a definition of reliability.

In their paper, Commenges and Jacqmin further study the relationship of the following propositions:

P1 $\rho_{ijj'} = \rho_i$ for all $j \neq j'$.

P2 $\tau_{ij} = \tau_i$ for all j.

P3 The model belongs to a class specified (with probability equal to one) by

$$\mathrm{E}\,(Y_{ij}|A_i) = k_{ij}\left[\lambda_i(A_i) + \eta_{ij}\right], \tag{3.3}$$
$$\mathrm{var}\,(Y_{ij}|A_i) = k_{ij}^2\left[\sigma_i^2(A_i) + \psi_{ij}(A_i)\right], \tag{3.4}$$

where k_{ij} and η_{ij} are deterministic quantities, $\mathrm{E}\,[\psi_{ij}(A_i)] = 0$, and $\psi_{ij}(A_i) > -\sigma_i^2(A_i)$.

P4 $\rho_i = \tau_i$.

Commenges and Jacqmin denote a model satisfying **P1**–**P4** a GICRE (general intraclass correlation random effect) model. In their paper, it is claimed that the propositions **P1**–**P3** are equivalent under the model (3.1). The following counterexamples should deny this claim.

Example 3.1 (Counterexample to **P2** ⇒ **P1**, **P3**).
Let $A_i \sim \mathrm{N}(0,1)$, $e_{ij} \sim \mathrm{N}(0,1)$, $i = 1,\ldots,n$, $j = 1,2,3$ *be mutually independent. Let*

$$Y_{ij} = c_j A_i^{2j-1} + e_{ij},$$

where constants c_j *are defined by*

$$c_j = 1/\sqrt{\mathrm{var}\,(A_i^{2j-1})},$$

so that $\mathrm{var}\,(c_j A_i^{2j-1}) = 1$. *For the Gaussian distribution we have*

$$\mathrm{E}\,A_i^{2k} = \frac{(2k-1)!}{2^{k-1}(k-1)!},$$
$$\mathrm{E}\,A_i^{2k-1} = 0.$$

Therefore, c_j *is more explicitly equal to*

$$c_j = \sqrt{\frac{2^{2k-2}(2k-2)!}{(4k-3)!}},$$

giving

$$c_1 = 1, \qquad c_2 = \frac{1}{\sqrt{15}}, \qquad c_3 = \frac{1}{3\sqrt{105}}.$$

Therefore we have

$$\rho_{i12} = \frac{\mathrm{cov}\,(c_1 A_i, c_2 A_i^3)}{2} = \frac{3}{4\sqrt{15}} \neq \frac{5}{2\sqrt{105}} = \frac{\mathrm{cov}\,(c_1 A_i, c_3 A_i^5)}{2} = \rho_{i13},$$

but

$$\tau_{ij} = \frac{\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)]}{\mathrm{var}\,(Y_{ij})} = \frac{\mathrm{var}\,(c_j A_i^{2j-1})}{\mathrm{var}\,(c_j A_i^{2j-1} + e_{ij})} = \frac{1}{2} = \tau_i.$$

The model does not belong to a class specified (with probability equal to one) by (3.3)–(3.4): For example, the conditional mean value

$$\mathrm{E}\,(Y_{ij}|A_i) = c_j A_i^{2j-1} \neq k_{ij}\,[\lambda_i(A_i) + \eta_{ij}]\,.$$

Example 3.2 (Counterexample to **P1** ⇒ **P2**, **P3**).
Let $A_i \sim \mathrm{N}(0,1)$, $e_{ij} \sim \mathrm{N}(0,1)$, $i = 1, \dots, n$, $j = 1, \dots, m$ *be mutually independent. Let*

$$Y_{ij} = A_i^{2j-1} + d_j A_i^{2j-1} e_{ij},$$

where constants d_j *will be specified later on. Then for all* i, j *it holds that*

$$\mathrm{E}\,(Y_{ij}|A_i) = A_i^{2j-1},$$
$$\mathrm{var}\,(Y_{ij}|A_i) = d_j^2 A_i^{2(2j-1)}.$$

Therefore,

$$\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)] = \mathrm{E}\,A_i^{2(2j-1)},$$
$$\mathrm{E}\,[\mathrm{var}\,(Y_{ij}|A_i)] = d_j^2 \mathrm{E}\,A_i^{2(2j-1)},$$

which cannot be written as in (3.3)–(3.4). Neither **P2** *holds in the case of* $d_j \neq d_{j'}$ *for some* $j \neq j'$, *since*

$$\tau_{ij} = \frac{\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)]}{\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)] + \mathrm{E}\,[\mathrm{var}\,(Y_{ij}|A_i)]} = \frac{1}{1 + d_j^2}.$$

For $j \neq j'$ *we have*

$$\varphi_{ijj'} = \mathrm{corr}[\mathrm{E}\,(Y_{ij}|A_i), \mathrm{E}\,(Y_{ij'}|A_i)] = \frac{\mathrm{cov}\,(A_i^{2j-1}, A_i^{2j'-1})}{\sqrt{\mathrm{var}\,(A_i^{2j-1})\mathrm{var}\,(A_i^{2j'-1})}}$$
$$= \frac{\mathrm{E}\,A_i^{2(j+j'-1)}}{\sqrt{\mathrm{E}\,A_i^{2(2j-1)}\mathrm{E}\,A_i^{2(2j'-1)}}} = \frac{(2j+2j'-3)!}{(j+j'-2)!}\sqrt{\frac{(2j-2)!}{(4j-3)!}}\sqrt{\frac{(2j'-2)!}{(4j'-3)!}},$$

and the correlation between Y_{ij} *and* $Y_{ij'}$ *can be written as (see Lemma 3.1)*

$$\rho_{ijj'} = \varphi_{ijj'}\sqrt{\tau_{ij}}\sqrt{\tau_{ij'}} = \varphi_{ijj'}\sqrt{\frac{1}{1+d_j}}\sqrt{\frac{1}{1+d_{j'}}}.$$

By the appropriate choice of constants d_j in our example, we can achieve the equality of $\rho_{ijj'}$ for all $j \neq j'$. To be more concrete, let us suppose $m = 3$. Then

$$\varphi_{i12} = \varphi_{i21} = \sqrt{\frac{3}{5}}, \qquad \varphi_{i13} = \varphi_{i31} = \sqrt{\frac{5}{21}}, \qquad \varphi_{i23} = \varphi_{i32} = \frac{\sqrt{7}}{3}.$$

When

$$d_1 = \sqrt{\frac{2}{7}}, \qquad d_2 = \frac{4}{\sqrt{5}}, \qquad d_3 = \sqrt{\frac{2}{3}},$$

we get

$$\tau_{i1} = \frac{7}{9}, \qquad \tau_{i2} = \frac{5}{21}, \qquad \tau_{i3} = \frac{3}{5},$$

but we get common correlation $\rho_{ijj'} = \rho_i = 1/3$.

As we have shown in Examples 3.1 and 3.2, propositions **P1**–**P3** cannot be considered generally equivalent. Let us now look closer at the relationship between $\rho_{ijj'}$ and τ_{ij}.

Lemma 3.1. *Suppose Y_{ij} obeys model (3.1). Then for all $j \neq j'$*

$$\rho_{ijj'} = \sqrt{\tau_{ij}}\sqrt{\tau_{ij'}}\,\mathrm{corr}[\mathrm{E}\,(Y_{ij}|A_i), \mathrm{E}\,(Y_{ij'}|A_i)] \leq \sqrt{\tau_{ij}}\sqrt{\tau_{ij'}}. \tag{3.5}$$

The equality in (3.5) holds for all $j \neq j'$ if for all $j \neq j'$

$$\mathrm{corr}[\mathrm{E}\,(Y_{ij}|A_i), \mathrm{E}\,(Y_{ij'}|A_i)] = 1, \tag{3.6}$$

that is, in the case when for all i, j the condition (3.3) with positive k_{ij} is true, which means that with probability equal to one, for some $k_{ij} > 0$, some η_{ij} and some function $\lambda_i(A_i)$, the conditional mean can be expressed as

$$\mathrm{E}\,(Y_{ij}|A_i) = k_{ij}\,[\lambda_i(A_i) + \eta_{ij}]\,.$$

Proof. Suppose $j \neq j'$. We start with a well known formula

$$\mathrm{cov}(Y_{ij}, Y_{ij'}) = \mathrm{cov}[\mathrm{E}\,(Y_{ij}|A_i), \mathrm{E}\,(Y_{ij'}|A_i)] + \mathrm{E}\,[\mathrm{cov}(Y_{ij}, Y_{ij'}|A_i)]. \tag{3.7}$$

The assumption of conditional independence implies

$$\mathrm{cov}(Y_{ij}, Y_{ij'}|A_i) = 0. \tag{3.8}$$

Therefore, the correlation coefficient is

$$\rho_{ijj'} = \frac{\mathrm{cov}[\mathrm{E}\,(Y_{ij}|A_i), \mathrm{E}\,(Y_{ij'}|A_i)]}{\sqrt{\mathrm{var}\,(Y_{ij})\mathrm{var}\,(Y_{ij'})}}.$$

From (3.2) we have

$$\mathrm{var}\,(Y_{ij}) = \frac{1}{\sqrt{\tau_{ij}}}\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)],$$
$$\mathrm{var}\,(Y_{ij'}) = \frac{1}{\sqrt{\tau_{ij'}}}\mathrm{var}\,[\mathrm{E}\,(Y_{ij'}|A_i)],$$

which together implies (3.5).
The correlation (3.6) is equal to one if for all $j = 1, \ldots, m$, the conditional means $\mathrm{E}\,(Y_{ij}|A_i)$ are with probability equal to one a linear function of one of them, for example if they are all a linear function of $\mathrm{E}\,(Y_{i1}|A_i) = \lambda_i(A_i)$.

Lemma 3.1 can be extended to the following theorem revising the Theorem of Commenges and Jacqmin:

Theorem 3.2 (Commenges and Jacqmin revised)**.** *Suppose that Y_{ij} for $i = 1, \ldots n$, $j = 1, \ldots m$, $m \geq 3$ obey the model (3.1). Moreover, with probability equal to one let for all i, j hold the assumption (3.3):*

$$\mathrm{E}\,(Y_{ij}|A_i) = k_{ij}\left[\lambda_i(A_i) + \eta_{ij}\right],$$

where $k_{ij} > 0$ and η_{ij} are deterministic quantities. Then the following propositions are equivalent:

P1 $\rho_{ijj'} = \rho_i$ *for all* $j \neq j'$.

P2 $\tau_{ij} = \tau_i$ *for all* j.

P3' *The model belongs to a class specified (with probability equal to one) by (3.4):*

$$\mathrm{var}\,(Y_{ij}|A_i) = k_{ij}^2\left[\sigma_i^2(A_i) + \psi_{ij}(A_i)\right],$$

where $\mathrm{E}\,[\psi_{ij}(A_i)] = 0$, *and* $\psi_{ij}(A_i) > -\sigma_i^2(A_i)$.

In addition, these propositions imply that

P4 $\rho_i = \tau_i$.

If moreover it holds

P5 $\lambda_i = \lambda$ *and* $\sigma_i^2 = \sigma^2$,

then also $\tau_i = \rho_i = \rho$.

Proof: Equivalence **P1** ⇔ **P2** is an easy consequence of (3.5):
Under (3.3) we have

$$\mathrm{corr}[\mathrm{E}\,(Y_{ij}|A_i), \mathrm{E}\,(Y_{ij'}|A_i)] = 1,$$

therefore

$$\rho_{ijj'} = \sqrt{\tau_{ij}}\sqrt{\tau_{ij'}}.$$

Now **P2** implies

$$\rho_{ijj'} = \sqrt{\tau_i}\sqrt{\tau_i} = \tau_i.$$

Vice-versa, if **P2** is not true, then there exist $j \neq j'$, such that $\tau_{ij} \neq \tau_{ij'}$. Having $m \geq 3$, there exist $k \neq j, j'$, and

$$\rho_{ijk} = \sqrt{\tau_{ij}}\sqrt{\tau_{ik}} \neq \sqrt{\tau_{ij'}}\sqrt{\tau_{ij'}} = \rho_{ij'k}.$$

Proof of **P3'** ⇒ **P1**, **P2**:

Let us suppose $j \neq j'$. Using (3.7), (3.8) and assumptions (3.3) and **P3'** (that is (3.4)), we can write

$$\begin{aligned}\mathrm{cov}(Y_{ij}, Y_{ij'}) &= \mathrm{cov}[\mathrm{E}\,(Y_{ij}|A_i), \mathrm{E}\,(Y_{ij'}|A_i)]\\ &= \mathrm{E}\,\{k_{ij}[\lambda_i(A_i)] - k_{ij}[\mathrm{E}\,(\lambda_i(A_i))]\} \cdot \{k_{ij'}[\lambda_i(A_i)] - k_{ij'}[\mathrm{E}\,(\lambda_i(A_i))]\}\\ &= k_{ij}k_{ij'}\mathrm{var}\,[\lambda_i(A_i)].\end{aligned}$$

Using

$$\mathrm{var}\,(Y_{ij}) = \mathrm{E}\,[\mathrm{var}\,(Y_{ij}|A_i)] + \mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)] \tag{3.9}$$

we have

$$\begin{aligned}\mathrm{var}\,(Y_{ij}) &= \mathrm{E}\,\{k_{ij}^2[\sigma_i^2(A_i) + \psi_{ij}(A_i)]\} + \mathrm{var}\,\{k_{ij}[\lambda_i(A_i) + \eta_{ij}]\}\\ &= k_{ij}^2\{\mathrm{E}\,[\sigma_i^2(A_i)] + \mathrm{var}\,[\lambda_i(A_i)]\},\\ \mathrm{var}\,(Y_{ij'}) &= k_{ij'}^2\{E[\sigma_i^2(A_i)] + \mathrm{var}\,[\lambda_i(A_i)]\}.\end{aligned}$$

Using $\mathrm{sign}(k_{ij}) = \mathrm{sign}(k_{ij'})$ we obtain

$$\begin{aligned}\rho_{ijj'} = \mathrm{corr}(Y_{ij}, Y_{ij'}) &= \frac{\mathrm{cov}(Y_{ij}, Y_{ij'})}{\sqrt{\mathrm{var}\,(Y_{ij})\mathrm{var}\,(Y_{ij'})}}\\ &= \frac{k_{ij}k_{ij'}\mathrm{var}\,[\lambda_i(A_i)]}{\sqrt{(k_{ij}k_{ij'})^2(\mathrm{E}\,[\sigma_i^2(A_i)] + \mathrm{var}\,[\lambda_i(A_i)])}} = \rho_i\end{aligned} \tag{3.10}$$

and finally also

$$\tau_{ij} = \frac{\mathrm{var}\;[\mathrm{E}\,(Y_{ij}|A_i)]}{\mathrm{var}\,(Y_{ij})} = \frac{\mathrm{var}\,[\lambda_i(A_i)]}{\mathrm{E}\,[\sigma_i^2(A_i)] + \mathrm{var}\,[\lambda_i(A_i)]} = \tau_i = \rho_i. \tag{3.11}$$

Proof of **P2** ⇒ **P3'**:
P2 together with (3.3) gives

$$\tau_{ij} = \frac{\mathrm{var}\,[\lambda_i(A_i)]}{\mathrm{var}\,[\lambda_i(A_i)] + \frac{\mathrm{E}\,[\mathrm{var}\,(Y_{ij}|A_i)]}{k_{ij}^2}}$$

which implies

$$\mathrm{E}\,[\mathrm{var}\,(Y_{ij}|A_i)] = k_{ij}^2 f_i(A_i),$$

where f_i is some function. Therefore

$$\mathrm{var}\,(Y_{ij}|A_i) = k_{ij}^2[\sigma_i^2(A_i) + \psi_{ij}(A_i)],$$

where E $[\psi_{ij}(A_i)] = 0$, and $\psi_{ij}(A_i) > -\sigma_i^2(A_i)$.

Proof of **P5**:
If moreover $\lambda_i = \lambda$ and $\sigma_i^2 = \sigma^2$, then neither $\mathrm{var}\,[\lambda(A_i)]$ nor $\mathrm{E}\,[\sigma^2(A_i)]$ do depend on i (recall that A_i have the same distributions). Thus (3.11) is equal to ρ. □

Remark 3.1. *Notice that implication* ***P1*** *⇒* ***P2*** *does not necessarily hold under (3.3) for $m = 2$, therefore the assumption $m \geq 3$ is needed.*

Remark 3.2. *In the original Theorem of Commenges and Jacqmin, authors do not require the assumption of $k_{ij} > 0$ for each i, j. We think, the assumption of the same sign of k_{ij}'s is needed to obtain the independence of j in (3.10). Nevertheless, the mentioned assumption is a reasonable one for our purposes, since it implies that all the items "grade in the same direction".*

Remark 3.3. *Moreover, we could for example set the constraint $\sum_{k=1}^m k_{ij}^2 = 1$ or the constraint $\sum_{k=1}^m k_{ij}^2 = m$ and multiply λ_i, σ_i^2, η_{ij} and ψ_{ij} by appropriate constants.*

Remark 3.4. *Assumption (3.3) is reasonable in our situation: if for some $j \neq j'$, $\mathrm{corr}[\mathrm{E}(Y_{ij}|A_i), \mathrm{E}(Y_{ij'}|A_i)] \neq 1$, then the j-th and the j'-th item of the educational test do not estimate exactly the same property. In some sense, the assumption (3.3) is equivalent to the assumption (1.6) of essential τ-equivalence in the classical model.*

We close this section by looking at the models discussed in previous sections in terms of Theorem 3.2.

Example 3.3 (Beta-binomial model)**.**
In the beta-binomial model (2.16) we have

$$\mathrm{E}(Y_{ij}|\pi_i) = \pi_i,$$
$$\mathrm{var}(Y_{ij}|\pi_i) = \pi_i(1-\pi_i),$$

therefore, the model satisfies (3.3) and (3.4) with $k_{ij} = 1$, $\lambda_i(\pi_i) = \pi_i$, $\eta_{ij} = 0$, $\sigma_i^2(\pi_i) = \pi_i(1-\pi_i)$ and $\psi_{ij}(\pi_i) = 0$. This agrees with the fact that

$$\rho_{ijj'} = \mathrm{corr}(Y_{ij}, Y_{ij'}) = \rho$$

and

$$\tau_{ij} = \frac{\mathrm{var}[\mathrm{E}(Y_{ij}|\pi_i)]}{\mathrm{var}(Y_{ij})} = \frac{\mathrm{var}(\pi_i)}{\mathrm{var}(Y_{ij})} = \frac{\mu(1-\mu)\rho}{\mu(1-\mu)} = \rho.$$

Example 3.4 (Model with additive item effect)**.**
Model (2.18) satisfies (3.3) since

$$\mathrm{E}(Y_{ij}|\pi_i) = \pi_i + b_j.$$

*Hence, by Theorem 3.2, the propositions **P1**–**P3'** are equivalent. The model does not satisfy **P3'**, since the conditional variance*

$$\mathrm{var}(Y_{ij}|\pi_i) = (\pi_i + b_j)(1 - (\pi_i + b_j))$$

cannot be written as in (3.4). This agrees with the fact that by Lemma 2.4 and by Lemma 2.2

$$\rho_{ijj'} = \rho\frac{1}{\sqrt{C_j C_{j'}}} \neq \rho_i$$

and

$$\tau_{ij} = \frac{\mathrm{var}[\mathrm{E}(Y_{ij}|\pi_i)]}{\mathrm{var}(Y_{ij})} = \frac{\sigma_\pi^2}{\mu(1-\mu) + b_j(1-2\mu-b_j)} \neq \tau_i$$

Remark: The probability that $\pi_i + b_j$ fall out of interval $(0,1)$ is neglected in this example.

To see when does the Rasch model satisfy (3.3) and **P1**–**P3'**, we need the following lemma:

Lemma 3.3. *Suppose that for some function $\lambda(x)$, for all x it holds that*

$$\frac{e^x}{\beta_j + e^x} = \frac{1}{\alpha_j}(\lambda(x) + \eta_j), \quad j = 1, \ldots, m, \tag{3.12}$$

where $m > 1$, $\alpha_j > 0$, $\beta_j > 0$. Then

$$\begin{aligned} \alpha_1 &= \cdots = \alpha_m, \\ \beta_1 &= \cdots = \beta_m, \\ \eta_1 &= \cdots = \eta_m. \end{aligned}$$

Proof: Expression (3.12) implies

$$\lambda(x) = \frac{\alpha_j e^x}{\beta_j + e^x} - \eta_j.$$

Suppose now any $1 \leq j \neq t \leq m$. Since (3.12) holds for all x, necessarily for all x

$$\frac{\alpha_j e^x}{\beta_j + e^x} - \eta_j = \frac{\alpha_t e^x}{\beta_t + e^x} - \eta_t.$$

The latter expression can be written as a polynomial in e^x. To hold for all x

$$A_0 + A_1 e^x + A_2 e^{2x} = 0,$$

A_0, A_1 and A_2 must be equal to zero. Therefore

$$\begin{aligned} A_0 &= \eta_t \beta_j \beta_t - \eta_j \beta_j \beta_t = 0 && \Rightarrow \eta_j = \eta_t = \eta, \\ A_2 &= \alpha_j - \alpha_t + \eta - \eta = 0 && \Rightarrow \alpha_j = \alpha_t = \alpha, \\ A_1 &= \beta_t \alpha - \beta_j \alpha - \eta(\beta_j + \beta_t) + \eta(\beta_j + \beta_t) = 0 && \Rightarrow \beta_j = \beta_t = \beta. \end{aligned}$$

□

Example 3.5 (Rasch model)**.**

*In the case of the constant item difficulties $b_1 = \cdots = b_m = b$, the Rasch model satisfies (3.3) and **P1**–**P3'** of Theorem 3.2 since*

$$\begin{aligned} \mathrm{E}\,(Y_{ij}|A_i) &= \frac{\exp(A_i + b)}{1 + \exp(A_i + b)} = \lambda(A_i) \\ \mathrm{var}\,(Y_{ij}|A_i) &= \frac{\exp(A_i + b)}{(1 + \exp(A_i + b))^2} = \sigma^2(A_i). \end{aligned}$$

In the case, when the item difficulties are not all the same, that is when $b_j \neq b'_j$ for some $j \neq j'$, the Rasch model does not satisfy (3.3): To satisfy (3.3), for all A_i it would be needed to hold

$$\mathrm{E}\,(Y_{ij}|A_i) = \frac{\exp(A_i + b_j)}{1 + \exp(A_i + b_j)} = \frac{e^{A_i}}{e^{-b_j} + e^{A_i}} = k_{ij}[\lambda_i(A_i) + \eta_{ij}]. \tag{3.13}$$

Nevertheless, by Lemma 3.3, (3.13) does not hold for unequal $\beta_j = e^{-b_j}$, hence the Rasch model with unequal difficulties does not satisfy (3.3).

In the next Chapter we will show that it does not satisfy either **P1** or **P2**, nevertheless that τ_{ij} are equal at least "approximately".

Chapter 4

Reliability of composite dichotomous measurements

Since no error term is assumed in models discussed in Chapter 2, the classical definition of reliability (1.2) is not applicable for these models. In this chapter, we propose a new, more general definition of reliability and we show the connection with the classical definition (1.2). Further we derive a formula for reliability in the model with general intraclass correlation (3.3)–(3.4), in the model with additive item effect (2.18), and in the Rasch model (2.2). In each model, we discuss the validity of the Spearman-Brown formula (1.8).

4.1 General definition of reliability

In Chapter 1, we defined the reliability by the ratio (1.2) comparing the variance of the measured property and the observed score variance. This definition is sensible for the assumed model (1.1), where the observed score is supposed to be a sum of the measured property and of an (independent) error term, and where the observed score variance is a sum of the variance of the measured property and of an error variance.

Nevertheless, the setting of model (1.1) is very restrictive. Therefore, to be more general, we may want to suppose only that measurement Y is somehow dependent on the true value of the measured property T :

$$Y \sim f(\bullet; T).$$

Such a general setting is supposed in Chapters 2 and 3 (see models (2.1) and (3.1)). When defining reliability in these modes, we can take an inspiration from coefficient τ_{ij} defined by (3.2) in Commenges and Jacqmin (1994). We propose to define the reliability more generally than in (1.2) by the ratio

$$R = \frac{\operatorname{var}\left[\operatorname{E}\left(Y|T\right)\right]}{\operatorname{var}\left(Y\right)}. \tag{4.1}$$

Similarly to the classical definition, there is the total observed variance in the denominator, and there is the part of the $\operatorname{var}(Y)$ due to variability of the measured property T in the numerator.

For the classical model (1.1), where

$$\mathrm{E}\,(Y|T) = T,$$

the new definition merges with the classical definition of reliability (1.2).

In the following sections, we derive formulas of reliability for models discussed in Chapter 2 and 3.

4.2 Reliability in the general intraclass correlation model

Suppose the model with general intraclass correlation, that is the model satisfying (3.3) and (3.4):

$$\begin{aligned}\mathrm{E}\,(Y_{ij}|A_i) &= k_{ij}\left[\lambda_i(A_i) + \eta_{ij}\right],\\ \mathrm{var}\,(Y_{ij}|A_i) &= k_{ij}^2\left[\sigma_i^2(A_i) + \psi_{ij}(A_i)\right],\end{aligned}$$

where $k_{ij} > 0$ and η_{ij} are given constants, $\mathrm{E}\,[\psi_{ij}(A_i)] = 0$, and $\psi_{ij}(A_i) > -\sigma_i^2(A_i)$. Let us moreover set the constraint of $\sum_{j=1}^m k_{ij}^2 = m$, discussed in Remark 3.3.

Then, the reliability of (every) single item can be written as

$$\begin{aligned}R_1 = \tau_{ij} &= \frac{\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)]}{\mathrm{var}\,[\mathrm{E}\,(Y_{ij}|A_i)] + \mathrm{E}\,[\mathrm{var}\,(Y_{ij}|A_i)]} = \frac{k_{ij}^2\mathrm{var}\,[\lambda_i(A_i)]}{k_{ij}^2\mathrm{var}\,[\lambda_i(A_i)] + k_{ij}^2\mathrm{E}\,[\sigma_i^2(A_i)]}\\ &= \frac{\mathrm{var}\,[\lambda_i(A_i)]}{\mathrm{var}\,[\lambda_i(A_i)] + \mathrm{E}\,[\sigma_i^2(A_i)]} = \tau_i = \rho_i = \mathrm{corr}(Y_{ij}, Y_{ij'}),\end{aligned}$$

and by Theorem 3.2 it merges with the correlation between two independent measurements of the same property Y_{ij}, $Y_{ij'}$, $j \neq j'$.

The reliability of the composite measurement can be expressed as

$$\begin{aligned}R_m &= \frac{\mathrm{var}\,[\mathrm{E}\,(Y_i|A_i)]}{\mathrm{var}\,[\mathrm{E}\,(Y_i|A_i)] + \mathrm{E}\,[\mathrm{var}\,(Y_i|A_i)]}\\ &= \frac{\mathrm{var}\,\left[\sum_{j=1}^m \mathrm{E}\,(Y_{ij}|A_i)\right]}{\mathrm{var}\,\left[\sum_{j=1}^m \mathrm{E}\,(Y_{ij}|A_i)\right] + \sum_{j=1}^m \mathrm{E}\,[\mathrm{var}\,(Y_{ij}|A_i)]}\\ &= \frac{\mathrm{var}\,\left\{\sum_{j=1}^m k_{ij}[\lambda_i(A_i) + \eta_{ij}]\right\}}{\mathrm{var}\,\left\{\sum_{j=1}^m k_{ij}[\lambda_i(A_i) + \eta_{ij}]\right\} + \sum_{j=1}^m k_{ij}^2\mathrm{E}\,[\sigma_i^2(A_i) + \psi_{ij}(A_i)]}\\ &= \frac{\left(\sum_{j=1}^m k_{ij}\right)^2 \mathrm{var}\,[\lambda_i(A_i)]}{\left(\sum_{j=1}^m k_{ij}\right)^2 \mathrm{var}\,[\lambda_i(A_i)] + m\mathrm{E}\,[\sigma_i^2(A_i)]}\\ &= \frac{\frac{\left(\sum_{j=1}^m k_{ij}\right)^2}{m} R_1}{1 + \left(\frac{\left(\sum_{j=1}^m k_{ij}\right)^2}{m} - 1\right) R_1}.\end{aligned} \tag{4.2}$$

Expression (4.2) is almost equivalent to the Spearman-Brown formula. It merges with Spearman-Brown formula if $k_{ij} = 1$ for all j. We can conclude that the assumptions (1.5)–(1.6) of τ-equivalence for model (1.1) correspond with assumption that with probability equal to one, the conditional mean and variance of Y_{ij} can be written as

$$\mathrm{E}\,(Y_{ij}|A_i) = \lambda_i(A_i) + \eta_{ij}, \tag{4.3}$$

$$\mathrm{var}\,(Y_{ij}|A_i) = \sigma_i^2(A_i) + \psi_{ij}(A_i), \tag{4.4}$$

where η_{ij} are given constants, $\mathrm{E}\,[\psi_{ij}(A_i)] = 0$, and $\psi_{ij}(A_i) > -\sigma_i^2(A_i)$.

Notice that if we were interested in the question for which combination of k_{ij} $j = 1, \ldots, m$ does for given λ_i and σ_i^2 the expression (4.2) reach its maximum, the solution of this simple problem of constrained optimization would be k_{ij} equal for all j.

4.3 Reliability in model with additive item effect

Using Lemma 2.4 we can easily derive formula for the reliability of the j-th item R_{1_j} and for the reliability of the total score R_m in model (2.18):

$$R_{1_j} = \tau_{ij} = \frac{\mathrm{var}\,\mathrm{E}\,(Y_{ij}|\pi_i)}{\mathrm{var}\,(Y_{ij})} = \frac{\sigma_\pi^2}{\mu(1-\mu) + b_j(1 - 2\mu - b_j)}. \tag{4.5}$$

Notice that the reliability R_1 in the model with no additive effect (that is when $b_j = 0$ for all $j = 1, \ldots, m$) equals to the correlation between Y_{ij} and $Y_{ij'}$ which is $\rho = \frac{\sigma_\pi^2}{\mu(1-\mu)}$.

The reliability R_m of the total score $Y_i = \sum_{j=1}^m Y_{ij}$ can be expressed as

$$\begin{aligned} R_m &= \frac{\mathrm{var}\,\mathrm{E}\,(Y_i|\pi_i)}{\mathrm{var}\,(Y_i)} = \frac{m^2\sigma_\pi^2}{m\mu(1-\mu)(1 + (m-1)\rho) - \kappa_b} \\ &= \frac{m\rho}{1 + (m-1)\rho - \frac{\kappa_b}{\mu(1-\mu)}}, \end{aligned} \tag{4.6}$$

where $\kappa_b = \frac{1}{m}\sum_{j=1}^m b_j^2$ describes the variability of items' difficulty. When $b_j = 0$ for all j, then (4.6) merges with the Spearman-Brown formula (1.8). Nevertheless, this is already the consequence of Section 4.2, since in such a case, the assumptions (4.3) and (4.4) are fulfilled.

We should recall that we neglect the small probability that $\pi_i + b_j$ fall out of the interval $(0, 1)$, and that we suppose b_j being "small numbers" and π_i being "not too close to 0 or 1." Therefore also $\frac{\kappa_b}{\mu(1-\mu)}$ is small and we do not face the situation when the denominator is zero or even negative.

Formula (4.6) implies that we attain greater reliability when the variability of items' difficulty is bigger. Nevertheless, as mentioned above, this difference is supposed to be small.

4.4 Reliability in the Rasch model

As follows from Example 3.5, in the case of equal item difficulties, the Rasch model (2.2) satisfies assumptions (4.3) and (4.4). As a consequence of Section 4.2, in such a case, the reliability R_m of the composite measurement Y_i obeys the Spearman-Brown formula

$$R_m = \frac{mR_1}{1+(m-1)R_1},$$

where the reliability of (every) single item is

$$R_1 = \frac{\operatorname{var}\left[\operatorname{E}(Y_{ij}|A_i)\right]}{\operatorname{var}(Y_{ij}} = \frac{\operatorname{var}\left[\lambda(A_i)\right]}{\operatorname{var}\left[\lambda(A_i)\right] + \operatorname{E}\left[\sigma^2(A_i)\right]} = \frac{\operatorname{var}\left(\frac{e^{A_i+b}}{1+e^{A_i+b}}\right)}{\operatorname{var}\left(\frac{e^{A_i+b}}{1+e^{A_i+b}}\right) + \operatorname{E}\left(\frac{e^{A_i+b}}{(1+e^{A_i+b})^2}\right)}$$

$$= \frac{\operatorname{E}\left(\frac{e^{A_i+b}}{1+e^{A_i+b}}\right)^2 - \left(\operatorname{E}\frac{e^{A_i+b}}{1+e^{A_i+b}}\right)^2}{\operatorname{E}\left(\frac{e^{A_i+b}}{1+e^{A_i+b}}\right)^2 - \left(\operatorname{E}\frac{e^{A_i+b}}{1+e^{A_i+b}}\right)^2 + \operatorname{E}\left(\frac{e^{A_i+b}}{(1+e^{A_i+b})^2}\right)} = \frac{C-D^2}{C-D^2+B}. \tag{4.7}$$

Unfortunately, integrals C, D and B in (4.7) cannot be evaluated explicitly, nevertheless their value can be approximated numerically.

In the case, when the item difficulties are not all equal, the reliabilities of items do differ, too:

$$R_{1_j} = \frac{C_{jj} - D_j^2}{C_{jj} - D_j^2 + B_j}, \tag{4.8}$$

where

$$B_j = \operatorname{E}\frac{e^{A+b_j}}{(1+e^{A+b_j})^2} = \int_{-\infty}^{\infty} \frac{e^{A+b_j}}{(1+e^{A+b_j})^2}\frac{1}{\sqrt{2\pi\sigma_A^2}}e^{-\frac{A^2}{2\sigma_A^2}}\,\mathrm{d}A,$$

$$D_j = \operatorname{E}\frac{e^{A+b_j}}{1+e^{A+b_j}} = \int_{-\infty}^{\infty} \frac{e^{A+b_j}}{1+e^{A+b_j}}\frac{1}{\sqrt{2\pi\sigma_A^2}}e^{-\frac{A^2}{2\sigma_A^2}}\,\mathrm{d}A$$

and

$$C_{jt} = \operatorname{E}\frac{e^{A+b_j}}{1+e^{A+b_j}}\frac{e^{A+b_t}}{1+e^{A+b_t}} = \int_{-\infty}^{\infty} \frac{e^{A+b_j}}{1+e^{A+b_j}}\frac{e^{A+b_t}}{1+e^{A+b_t}}\frac{1}{\sqrt{2\pi\sigma_A^2}}e^{-\frac{A^2}{2\sigma_A^2}}\,\mathrm{d}A.$$

In the case of unequal item difficulties, the reliability R_m of the composite measurement Y_i can be expressed using the integrals mentioned above

$$R_m = \frac{\sum_{j=1}^m \sum_{t=1}^m (C_{jt} - D_jD_t)}{\sum_{j=1}^m \sum_{t=1}^m (C_{jt} - D_jD_t) + \sum_{j=1}^m B_j}. \tag{4.9}$$

Table 4.1 shows the values of the reliability for some numbers of items m and some variabilities of student abilities σ_A, when the equidistantly distributed item difficulties between -0.1 and 0.1 of length m are chosen.

The values were calculated using function `integrate` in software `R`, using multiple of ±25 of the variability σ_A as the limits of integration. The maximum absolute

Table 4.1: Reliability in the Rasch model for different number of items

Number	Variability of abilities σ_A						
of items	0.01	0.1	0.2	0.5	0.9	2.5	10
m=3	0.00008	0.00741	0.02881	0.15047	0.34335	0.73121	0.94152
m=11	0.00028	0.02667	0.09814	0.39386	0.65731	0.90890	0.98335
m=20	0.00050	0.04747	0.16519	0.54160	0.77717	0.94775	0.99077
m=50	0.00125	0.11078	0.33098	0.74709	0.89711	0.97843	0.99629
m=100	0.00249	0.19947	0.49735	0.85524	0.94577	0.98910	0.99814

error reached in integrations for $m = 3$, $m = 11$, and $m = 20$ was less than 0.000025, for $m = 50$ and $m = 100$ it was less than 0.00013.

Let us now look "how far" the numerical values of Table 4.1 are from the values achieved by Spearman-Brown formula (1.9). Let us set for example $m_1 = 11$ and take the values of the second line of Table 4.1 as R_{m_1}. Then, the Spearman-Brown formula would give us the following values of reliabilities for $m = 3$, 20, 50, and 100 :

Table 4.2: Spearman-Brown formula used for $m_1 = 11$.

Number	Variability of abilities σ_A						
of items	0.01	0.1	0.2	0.5	0.9	2.5	10
SB R_3	0.00008	0.00742	0.02882	0.15054	0.34345	0.73125	0.94153
m=11	**0.00028**	**0.02667**	**0.09814**	**0.39386**	**0.65731**	**0.90890**	**0.98335**
SB R_{20}	0.00050	0.04746	0.16518	0.54159	0.77716	0.94775	0.99077
SB R_{50}	0.00125	0.11077	0.33095	0.74707	0.89710	0.97843	0.99629
SB R_{100}	0.00249	0.19944	0.49731	0.85522	0.94576	0.98910	0.99814

The numerical values in tables 4.1 and 4.2 are very similar and it is a matter of question, to what extent the differences are due to integration error. In the following, we will try to show that the Spearman-Brown formula holds in the Rasch model with unequal difficulties at least approximately.

Let us assume $\sum_j b_j = 0$, and apply the Taylor approximation:

$$\begin{aligned}
B_j &= \mathrm{E}\frac{e^{A+b_j}}{(1+e^{A+b_j})^2} \approx \mathrm{E}\frac{e^A}{(1+e^A)^2} + b_j\mathrm{E}\frac{e^A(1-e^A)}{(1+e^A)^3} = B + b_j\mathrm{E}\frac{e^A(1-e^A)}{(1+e^A)^3}, \\
D_j &= \mathrm{E}\frac{e^{A+b_j}}{1+e^{A+b_j}} \approx \mathrm{E}\frac{e^A}{1+e^A} + b_j\mathrm{E}\frac{e^A}{(1+e^A)^2} = D + b_j\mathrm{E}\frac{e^A}{(1+e^A)^2}, \\
C_{jt} &= \mathrm{E}\frac{e^{A+b_j}}{1+e^{A+b_j}}\frac{e^{A+b_t}}{1+e^{A+b_t}} \approx \mathrm{E}\frac{e^{2A}}{(1+e^A)^2} + (b_j+b_t)\mathrm{E}\frac{2e^{2A}}{1+e^A} = \\
&= C + (b_j+b_t)\mathrm{E}\frac{2e^{2A}}{1+e^A}. \qquad (4.10)
\end{aligned}$$

Therefore the reliability of the composite measurement is approximately

$$R_m = \frac{m^2(C-D^2)}{m^2(C-D^2)+mB}.$$

When assuming moreover that b_j's are close enough to zero, we have

$$R_1 \approx \frac{C - D^2}{C - D^2 + B},$$

which gives an approximate validity of the Spearman-Brown formula.

As for conclusion of this chapter, we have shown that the reliability of composite dichotomous measurement can be defined even if the item reliabilities (and the intraclass correlations) are not the same for all items (all pairs of items). In such a case, of course, we can hardly talk about validity of the Spearman-Brown formula. Nevertheless, if the item difficulties differ only slightly, the Spearman-Brown formula is valid at least approximately.

Chapter 5

Estimation of reliability of composite dichotomous measurements

5.1 Cronbach's alpha and *logistic alpha*

As we have already mentioned in Section 1.3, Cronbach's alpha was in fact designed as a generalization of so called Kuder-Richardson formula 20 for dichotomous scoring, proposed already in Kuder and Richardson (1937):

$$\hat{\alpha} = \frac{m}{m-1} \frac{s^2 - \sum_{j=1}^{m} p_j(1-p_j)}{s^2}, \tag{5.1}$$

where p_j is a relative frequency of correct answers to the j-th item and

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

is a sample estimate of variance of total scores. One can easily see that (5.1) can be obtained when computing the sample estimate of Cronbach's alpha (1.10) in case of dichotomous scoring, where $\hat{\mathrm{E}} \sum_i Y_{ij}/n = p_j$ and $\widehat{\mathrm{var}} \sum_i Y_{ij}/n = p_j(1-p_j)$.

Nevertheless, with dichotomous items, the assumptions of analysis of variance are violated. The scores cannot be assumed to have normal distribution and, moreover, the variance is dependent on the mean value. Therefore it is a matter of question to what extent is this estimate appropriate at all. To answer this question, we should investigate the properties of Cronbach's alpha in the models described in Chapter 2.

To modify Cronbach's alpha for the case of binary outcomes, the following idea (see Zvára (2002)) can cross our mind: while the F-statistic in (1.25) is best suited for normally distributed variables, we should replace it by an analogous statistic appropriate for dichotomous data.

Testing the hypothesis $H_0 : \mathrm{var}\,(T) = 0$ is equal to testing the submodel B where the score Y_{ij} depends only on the test item (and does not depend on the student's ability) against the model A+B where the score Y_{ij} depends on the student and on the test item. In the fixed-effect model of logistic regression, the appropriate statistic is the difference of deviances in the submodel and in the model

$$X^2 = D(B) - D(A + B), \tag{5.2}$$

where deviance D is defined as a function of the difference of the log-likelihood for the model and for the saturated model (for details see for example Agresti (2002), pp.139). Statistics (5.2) has under the null hypothesis asymptotically (for n fixed and m approaching infinity) the $\chi^2(n-1)$ distribution. Therefore, the proposed estimate is

$$\hat{\alpha}_{log} = 1 - \frac{n-1}{X^2}. \tag{5.3}$$

In the next section, we demonstrate the properties of Cronbach's alpha and of the logistic alpha (5.3) in the Rasch model and in the modified Rasch model.

5.2 Simulations

First, we supposed that the data come from the Rasch model. We studied the case of $n = 20$ students and $m = 11$ items, which corresponds to common situation in high-school classes. Besides, the number of $n = 30$ and of $n = 50$ students, and the number of $m = 20$ and $m = 50$ items was studied. The item difficulties were always taken equidistant between -0.1 and 0.1. In each case, the number of the 55 values of σ_A were chosen so that the resulting 55 reliabilities would cover the interval $\langle 0, 1 \rangle$.

For each of the five combinations of number of students and number of items (figures 5.1 – 5.5) and for each of 55 values of σ_A (55 points in the figure), the true reliability was computed via formula (4.9). Further, the following procedure was repeated 500-times for each point:

1. The set of n student abilities A_i was generated from the $\mathrm{N}(0, \sigma_A)$ distribution
2. For each of n abilities A_i, the m scores on the test items were generated from the Rasch model (2.2)
3. The sample estimate of the Cronbach's alpha (1.25) and the logistic alpha (5.3) was computed from the data

For each of 500 sample estimates of Cronbach's alpha and logistic alpha, their average value and sample variance were computed, and finally, the bias and mean squared error (MSE) were displayed.

Already the first Figure 5.1 gives an impression that the new estimate gives better results (smaller bias and mean squared error), except for the case of the true reliability value close to 1. When looking at Figures 5.2 and 5.3, we can see that the properties of the new estimate are even better. We can conclude that the new estimate can estimate the reliability of the composite dichotomous measurement better than the classical estimate based on Cronbach's alpha, especially in the situation,

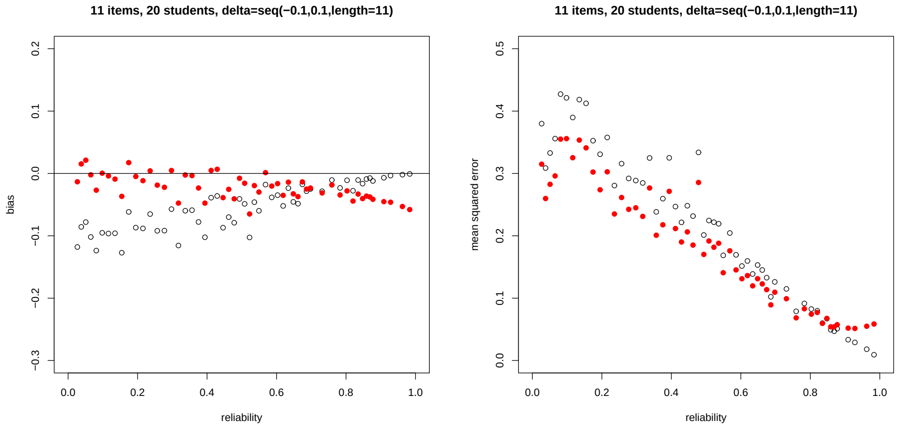

Figure 5.1: Bias and MSE for classical (empty circles) and logistic (solid circles) estimator of reliability. Number of students 20, number of items 11.

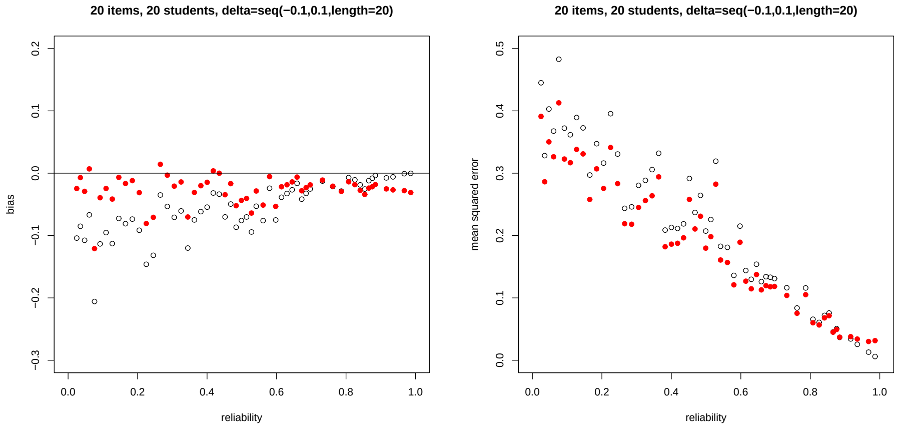

Figure 5.2: Bias and MSE for classical (empty circles) and logistic (solid circles) estimator of reliability. Number of students 20, number of items 20.

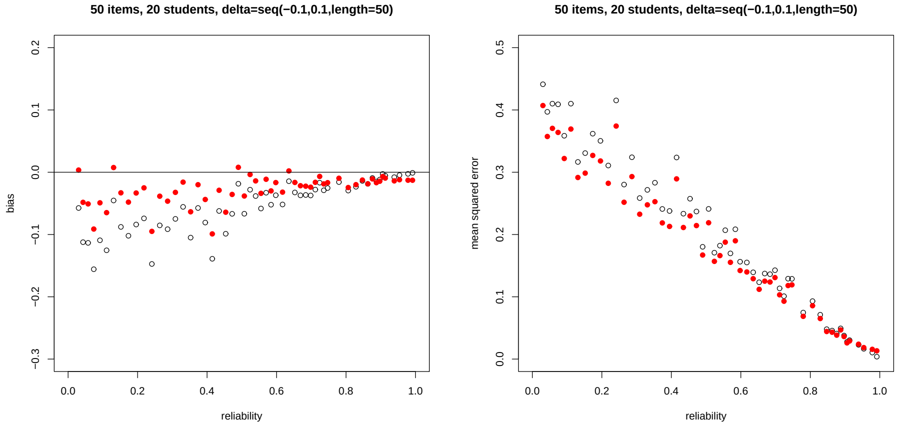

Figure 5.3: Bias and MSE for classical (empty circles) and logistic (solid circles) estimator of reliability. Number of students 20, number of items 50.

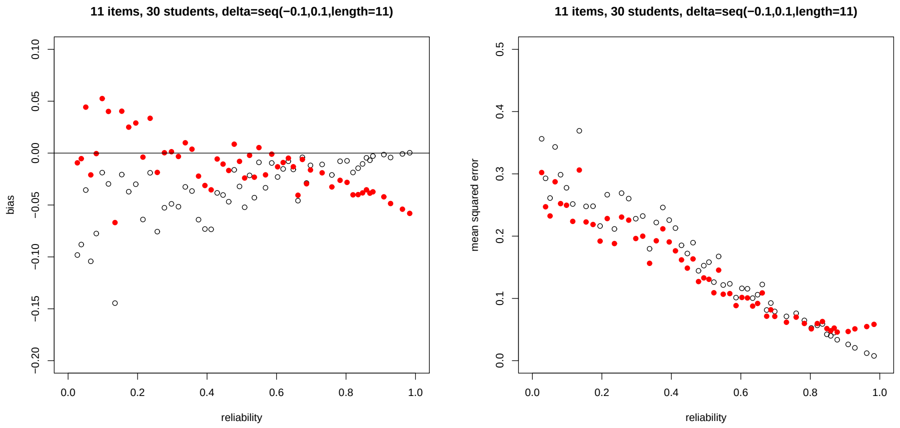

Figure 5.4: Bias and MSE for classical (empty circles) and logistic (solid circles) estimator of reliability. Number of students 30, number of items 11.

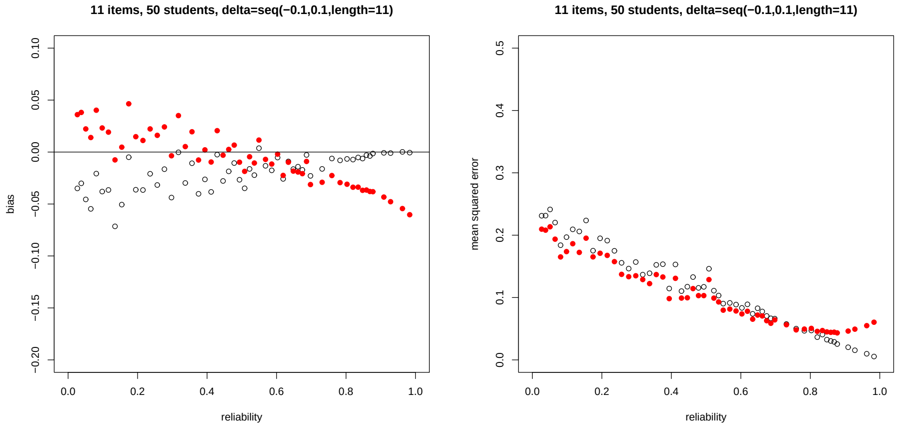

Figure 5.5: Bias and MSE for classical (empty circles) and logistic (solid circles) estimator of reliability. Number of students 50, number of items 11.

when the number of items is high, and when the true reliability is not too close to one. On the other hand, in the case of high number of students relatively to the number of items (Figures 5.4 and 5.5), the results of the new estimate are a bit worse for the true reliability close to one.

Similar results are obtained, when the data are supposed to come from the modified Rasch model (2.19). In Figure 5.6, the case of 20 students and 11 items is displayed. The trend is similar to Figure 5.1: The new estimate gives better results for true reliabilities not too close to 1. Higher bias (for both Cronbach and logistic alpha) for the true reliabilities close to zero is connected with the fact that in this case, the probability $\pi_i + b_j$ often exceeds the interval $(0, 1)$.

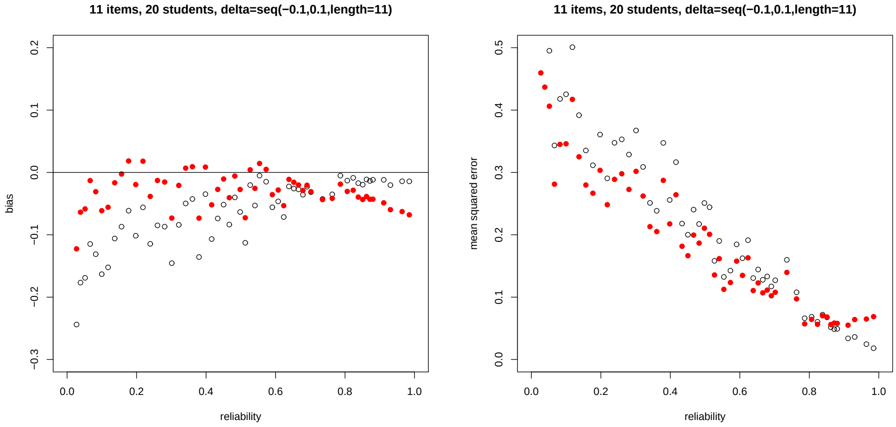

Figure 5.6: Simulation from the modified Rasch model: Bias and MSE for classical (empty circles) and logistic (solid circles) estimator of reliability. Number of students 20, number of items 11.

Conclusion

This work was concerned with the reliability of composite dichotomous measurements. Since the classical concept of reliability is not appropriate for the dichotomous data, a more general definition of reliability was proposed, of which the classical definition was shown to be a special case.

The proposed definition was motivated by coefficient τ_{ij} introduced in Commenges and Jacqmin (1994), where it is claimed to be in some sense equivalent to the intraclass correlation coefficient $\rho_{ijj'}$, which is known to merge with reliability for common correlation models. In Chapter 3 we revised the paper of Commenges and Jacqmin and we have shown that stronger assumptions are needed for the claimed equivalence to hold.

Since the discussed coefficient τ is of similar nature as the reliability (that is it compares the variability due to the measured property with the total variability of the measurement), we proposed its use as a more general definition of reliability. Further, we derived formulas of reliability in different models appropriate for dichotomous data. Even in the case when the reliabilities of items are not all equal, the reliability of the composite measurement can still be defined. We have shown that when the item difficulties differ only slightly, the relationship between the reliability of a single item and the reliability of the composite measurement obeys "at least approximately" the Spearman-Brown formula.

In the last chapter, a new estimate of reliability was proposed, which could be more appropriate in the case of binary data than the classical estimate based on Cronbach's alpha. Via simulations in the Rasch model, we have shown that the new estimate tends to give better results (smaller bias and mean squared error) than the classical estimate. Nevertheless, the new estimate gives inferior results when the true value of reliability is close to one or when the number of students is too high (when compared to the number of items). Further work should contain the study of the theoretical properties of the classical and the new estimate in the Rasch model. This could also lead to improvement of the proposed estimate for the mentioned critical situations.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis.* Wiley, New Jersey, USA.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society*, 32(2):283–301.

Bravo, G. and Potvin, L. (1991). Estimating the reliability of continuous measures with Cronbach's alpha or the intraclass correlation coefficient: Toward the integration of two traditions. *J. Clin. Epidemiol.*, 44:381–390.

Christmann, A. and Van Aelst, S. (2006). Robust estimation of Cronbach's alpha. *Journal of Multivariate Analysis*, 97:1660–1674.

Commenges, D. and Jacqmin, H. (1994). The intraclass correlation coefficient distribution-free definition and test. *Biometrics*, 50:517–526.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Eggen, T. J. H. M. (2000). On the loss of information in conditional maximum likelihood estimation of the item parameters. *Psychometrika*, 65:337–362.

Guttman, L. A. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 30:357–370.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, 6:153–160.

Kuder, G. and Richardson, M. (1937). The theory of estimation of test reliability. *Psychometrika*, 2:151–160.

Martinková, P. (2006). Reliability in the Rasch model. In Hakl, F., editor, *Proceedings of the XI. PhD. Conference of the ICS ASCR*, pages 64–71. Matfyzpress, Prague.

Martinková, P. and Zvára, K. (2007). Reliability in the Rasch model. *Kybernetika.* Submitted.

Martinková, P., Zvára, K., Zvárová, J., and Zvára, K. (2006). The new features of the ExaMe evaluation system and reliability of its fixed tests. *Methods of Information in Medicine*, 45:310–315.

Neyman, J. and Scott, E. L. (1948). Consistent estimators based on partially consistent observations. *Econometrica*, 16(1):1–32.

Novick, M. R. and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurement. *Psychometrika*, 32:1–13.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* The Danish Institute of Educational Research.

Rexová, P. (2003). *Reliability of measurements [Spolehlivost měření, in Czech]. Diploma thesis.* Charles University in Prague.

Rexová, P. (2004a). Item analysis of educational tests. In Hakl, F., editor, *Proceedings of the IX. PhD. Conference of the ICS ASCR*, pages 107–112. Matfyzpress, Prague.

Rexová, P. (2004b). Item analysis of educational tests. In Šafránková, J., editor, *WDS'04 Proceedings of Contributed Papers: Part I - Mathematics and Computer Sciences*, pages 77–83. Matfyzpress, Prague.

Ridout, M. S., Demetrio, C. G. B., and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics*, 55:137–148.

van der Linden, W. J. and Hambleton, R. K., editors (1997). *Handbook of item response theory.* Springer-Verlag, New York.

Wilcox, R. R. (1992). Robust generalizations of classical test reliability and Cronbach's alpha. *British Journal of Mathematical and Statistical Psychology*, 45:239–254.

Zou, G. and Donner, A. (2004). Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics*, 60:807–811.

Zvára, K. (2002). Measuring of reliability: Beware of Cronbach. [Měření reliability aneb bacha na Cronbacha, in Czech]. *Information bulletin of the Czech Statistical Society*, 12:13–20.