

CONTRIBUTIONS TO THE THEORY AND PRACTICE OF  
COMPUTERIZED ADAPTIVE TESTING

Theo J.H.M. Eggen

## Samenstelling promotiecommissie

Voorzitter/ secretaris	prof.dr. B.E. van Vucht-Tijssen
Promotor	prof.dr. N.D. Verhelst
Leden	prof.dr. C.A.W. Glas (Universiteit Twente)
	prof.dr. H. Hoijtink (Universiteit Utrecht)
	prof.dr. W.J. van der Linden (Universiteit Twente)
	prof.dr. K. Sijtsma (Universiteit van Tilburg)
	dr. H.H.F.M. Verstralen (Cito, Arnhem)

ISBN 90–5834-056-2

Omslag: Roel Ottow / Harold Kainama

Druk: Print Partners Ipskamp B.V., Enschede

© Copyright 2004, Theo J.H.M. Eggen, Citogroep Arnhem, NL

CONTRIBUTIONS TO THE THEORY AND PRACTICE OF  
COMPUTERIZED ADAPTIVE TESTING

PROEFSCHRIFT

ter verkrijging van  
de graad van doctor aan de Universiteit Twente,  
op gezag van de rector magnificus,  
prof.dr. F.A. van Vught,  
volgens het besluit van het College voor Promoties  
in het openbaar te verdedigen  
op vrijdag 12 maart 2004 om 15.00 uur

door

Theodorus Johannes Hendrikus Maria Eggen  
geboren op 4 April 1953  
te Babberich

Dit proefschrift is goedgekeurd door de promotor prof.dr. N.D. Verhelst.



## Voorwoord

Computergestuurd adaptief toetsen is het onderwerp van dit proefschrift. Bij deze vorm van toetsen wordt de moeilijkheid van de toets tijdens het toetsen constant aangepast aan het niveau van de kandidaat. Dit proefschrift bevat bijdragen die ik in het onderzoeksproject adaptief toetsen van het Cito heb kunnen maken.

Een proefschrift voltooien gaat alleen met de hulp van velen. Op deze plaats wil ik iedereen bedanken die heeft bijgedragen aan de totstandkoming. Enkelen wil ik daarbij in het bijzonder noemen.

Mijn promotor, Norman Verhelst, ben ik zeer erkentelijk voor zijn verhelderende discussies, zijn ideeën, zijn kritiek en zijn apl programma's. Mijn promotie-commissie dank ik voor de bereidheid het manuscript te lezen en te beoordelen.

In het onderzoeksproject adaptief toetsen en bij het ontwikkelen van adaptieve toetsen werk ik op prettige wijze samen met veel collega's. Van Gerard Straetmans, Norman Verhelst en Angela Verschoor zijn de bijdragen zo groot dat ze co-auteur zijn van een hoofdstuk.

Mijn werkgever, het Cito, wil ik danken voor het geven van faciliteiten voor het schrijven en het produceren van het proefschrift. Piet Sanders, hoofd van de afdeling Psychometrisch Onderzoek en Kenniscentrum, bedank ik voor zijn aansporingen om het af te maken en voor zijn bemiddeling die het Engels van mijn teksten een stuk beter maakten. Mijn collega's van de afdeling dank ik voor de prettige werksfeer en de door hen gegeven gevraagde en ongevraagde steun.

Het proefschrift bestaat uit een inleidend hoofdstuk, waarin de basiselementen van een computergestuurde adaptieve toets worden besproken en de zes overige hoofdstukken worden geïntroduceerd. Alle hoofdstukken zijn zelfstandig leesbaar. Het boek wordt afgesloten met een samenvatting in het Nederlands.



## Stellingen

behorende bij het proefschrift van Theo J.H.M. Eggen: *Contributions to the theory and practice of computerized adaptive testing*, maart 2004.

1. Moderne computergestuurde adaptieve toetsen meten optimaal binnen praktisch gewenste randvoorwaarden, waardoor het bij gegeven meetnauwkeurigheid de goedkoopste manier van computergestuurd toetsen is.
2. De grote kans op een specificatiefout in het model bij toepassing van de marginale maximale aannemelijkheidsmethode (MML) bij het schatten van de itemparameters weegt niet op tegen het geringe informatieverlies dat men lijdt bij toepassing van de conditionele maximale aannemelijkheidsmethode (CML).
3. Hoewel softwarepakketten voor de kalibratie van itembanken gegevens verzameld in onvolledige designs kunnen analyseren, houden ze geen rekening met het stochastisch karakter van de ontbrekende gegevens, waardoor de analyse-resultaten waardeloos kunnen zijn.
4. Door toepassing van moderne dataverzameling-, dataopslag- en communicatie-technieken, in combinatie met het gebruik van empirisch Bayesiaanse schattingsmethoden, wordt het ontbreken van noodzakelijke pretest-gegevens over opgaven slechts een zeer tijdelijk probleem.
5. De praktijk van het steekproef trekken in het Nederlandse onderwijs, leidt er noodzakelijkwijs toe om voor de kalibratie van itembanken een sterke voorkeur te hebben voor de conditionele maximale aannemelijkheidsmethode om itemparameters te schatten.
6. De stabiliteit van itemparameters in (computergestuurde) (adaptieve) toetsen verdient meer aandacht dan ze doorgaans krijgt.
7. Naarmate het belang van de beslissing op grond van een toets toeneemt, dient men voorzichter om te gaan met individualisering en flexibilisering. Het is daarom onverantwoord om eindexamens zonder empirische evidentie van de meetkwaliteit te gebruiken.
8. De universiteiten, wellicht in samenwerking met het Cito, zouden een gezamenlijke inspanning moeten plegen om het vak van psychometrie, of zo men wil de onderwijsmeetkunde, voor studenten aantrekkelijk te maken.
9. Het Cito is groot geworden en kan alleen groot blijven door kwalitatief hoogwaardige toetsen in een professionele werkomgeving te ontwikkelen.
10. Na de invoering van rookzuilen op perrons van spoorwegstations, lijkt de tijd ook rijp voor belplaatsen, geurplekken, vloekhoeken en praatpalen.

## Table of contents

<b>1. Introduction and overview .....</b>	<b>1</b>
1.1 Introduction .....	3
1.2 Overview .....	9
1.3 References .....	11
<b>2. On the loss of information in conditional maximum likelihood     estimation of item parameters .....</b>	<b>13</b>
2.1 Introduction .....	15
2.1.1 Estimation of item parameters in the Rasch model .....	15
2.1.2 Information and efficiency .....	17
2.2 Notation and terminology .....	19
2.3 The F-information: definition and basic properties .....	20
2.4 A scalar measure of information .....	24
2.5 F-information in separable models .....	25
2.5.1 Properties of the efficient score statistics .....	26
2.5.2 Theorems on the F-information in separable models .....	27
2.6 Checking the conditions for no loss of information using CML in two Rasch models .....	32
2.6.1 The Rasch Poisson counts model .....	32
2.6.2 The Rasch model for dichotomously scored items .....	34
2.7 F-information in the dichotomous Rasch model: comparing JML and CML .....	35
2.7.1 Comparison of information efficiency in JML and CML estimation .....	41
2.8 F-information in the dichotomous Rasch model: comparing MML and CML .....	44
2.8.1 F-information in $p_m$ .....	45
2.8.2 F-information in $f(x t)$ .....	49
2.8.3 Comparison of information efficiency in CML and MML estimation .....	51
2.9 Conclusion .....	54
2.10 References .....	56
Appendix chapter 2 .....	59

<b>3. Loss of information in estimating item parameters</b>	
<b>in incomplete designs</b>	61
3.1 Introduction	63
3.2 F-information and the Rasch model	63
3.3 Normalization, information, and the determinant	68
3.3.1 The influence of the normalization on the information matrices	69
3.4 F-information in incomplete designs	74
3.5 The information comparison in incomplete designs	78
3.6 Examples of comparing the efficiency in incomplete designs	81
3.6.1 The designs	82
3.6.2 Results for an observed response as unit of cost	85
3.6.3 Results for a test taker as unit of cost	91
3.7 Conclusion	93
3.8 References	95
Appendix chapter 3	96
 <b>4. Item calibration in incomplete testing designs</b>	 97
4.1 Introduction	99
4.2 Item response theory	100
4.2.1 Conditional maximum likelihood estimation	101
4.2.2 Marginal maximum likelihood estimation	102
4.3 Inference and missing data	103
4.4 Incomplete calibration designs	107
4.4.1 Random incomplete designs	108
4.4.2 Multistage testing designs	109
4.4.3 Targeted testing designs	110
4.5 Item calibration and missing data	112
4.5.1 The marginal model and missing data	113
4.5.2 The conditional model and missing data	122
4.6 Conclusion	131
4.7 References	133

<b>5. Computerized adaptive testing for classifying</b>	
<b>examinees into three categories</b> .....	135
5.1 Introduction .....	137
5.2 Context .....	139
5.2.1 The mathematics item bank .....	139
5.3 Research questions .....	142
5.4 Statistical computation procedures .....	142
5.4.1 Statistical estimation in the testing algorithm .....	143
5.4.2 Statistical testing in the testing algorithm .....	144
5.5 Item selection .....	147
5.5.1 Starting procedure .....	147
5.5.2 Item selection procedures .....	147
5.6 Design of the simulation studies .....	149
5.7 Results of the simulation studies .....	152
5.7.1 Measurement accuracy with statistical estimation .....	152
5.7.2 The algorithms in the conditions of the placement test .....	153
5.8 Discussion .....	161
5.9 References .....	164
 <b>6. Item selection in adaptive testing with the sequential</b>	
<b>probability ratio test</b> .....	167
6.1 Introduction .....	169
6.2 Sequential testing in the testing algorithm .....	170
6.2.1 Classification in two categories .....	170
6.2.2 Classification in three categories .....	172
6.3 Item selection .....	174
6.3.1 Fisher information .....	175
6.3.2 Kullback-Leibler information .....	177
6.4 Comparison of item selection procedures .....	180
6.4.1 Method .....	181
6.4.2 Results .....	183
6.5 Discussion .....	189
6.6 References .....	192

<b>7. Optimal testing with easy or difficult items in computerized</b>	
<b>adaptive testing</b> .....	195
7.1 Introduction .....	197
7.2 Item selection in CAT .....	199
7.3 Item selection on the basis of success probability .....	201
7.3.1 Performance of item selection based on nearest p-point .....	202
7.4 Alternative method for selecting with higher or lower success probabilities .....	208
7.4.1 Performance of item selection based on selection at a shifted ability level .....	211
7.4.2 Some properties of selection at the shifted ability level .....	213
7.5 Discussion .....	217
7.6 References .....	219
<b>Samenvatting</b> .....	221

# Chapter 1

## Introduction and overview





## **1.1 Introduction**

In computerized adaptive tests (CATs) the construction and administration of the test is computerized and individualized. For every test taker a different test is constructed by selecting items from an item bank tailored to the ability of the test taker as demonstrated by the responses given thus far. So, in principle, each test taker is administered a different test whose composition is optimized for the person. The main motive for computerized adaptive testing is efficient measurement. It has been shown that CATs need less items to measure the ability of the test taker with the same precision. Since the publication of the basic ideas on modern computerized adaptive testing in Lord (1970), the educational and psychometric community has produced numerous articles and books on this subject. Recently, two books have been published which present an overview. The volume, edited by Wainer (2000), gives the historical development and the basics of computerized adaptive testing. It gives a description of possibilities for building, maintaining and using CATs. The volume, edited by Van der Linden & Glas (2000), is a compilation of recent psychometric research on CATs.

In this chapter, a general overview of the basic aspects of computerized adaptive testing will be presented first. Thereafter, the topics addressed in the subsequent chapters of this dissertation will be introduced.

### Prerequisites and limitations of computerized adaptive testing

Computerized testing trivially assumes the availability of an automated environment, hardware and software, which performs the testing according to acceptable standards. Although the developments in information and communication technology in recent years show an exponential growth in possibilities and performance, this does not imply that there are no limitations for successfully applying computerized adaptive testing. Most restraints apply to computer based testing in general, but there are aspects of Cats which put extra demands on the available computer capabilities. The main points are that in a CAT during testing a complete item bank should be directly accessible and that computational procedures for ability estimation and item selection must be

carried out in real time. More general limitations in computerized testing are imposed on the item format or response format. Although there are ample possibilities for the type of items or stimuli which can be presented to candidates, there are still main limitations in the permissible responses of the candidates. This is due to the fact that these responses need automatic scoring. That is why items in computerized tests often are of a choice or matching type, and in case the candidate can give their own response the response formats are limited. For example, a few words to complete a sentence or the results of a computation. Another limitation of adaptive testing related to the admissible response format, is that most adaptive testing algorithms assume that the responses to items are dichotomously scored, correct or incorrect. Although some research results are available for adaptive testing with polytomously scored items (Dodd, De Ayala & Koch, 1995) and Van Rijn, Eggen, Hemker & Sanders, 2002), practical applications are limited; the main reason is the practical difficulty of scoring polytomous items automatically. For this reason, the research reported on computerized adaptive testing in this thesis is limited to dichotomously scored items.

### Item bank and item response theory

CATs presuppose the availability of an item bank. An item bank is a structured collection of items which are constructed to measure a well defined ability of persons. The item bank contains besides (a link to) the items themselves, various characteristics of each item. The characteristics of the items may concern a content classification or administrative data on the items, but the most important characteristics are the psychometric characteristics of the items. The psychometric characteristics of the items are the results of the application of a test theory. In such a theory a relation is specified between the non-observable ability  $\theta$  which is measured and the observable score on the measurement instrument. CATs make use of item response theory (IRT) (Van der Linden & Hambleton, 1996). In the case of dichotomously scored items, the most popular item response models are of the logistic type. In these models, a specification is

given of the relation between the ability,  $\theta$ , of a person and the probability of correctly answering item  $i$ ,  $X_i = 1$ . The exact relationship is determined by characteristics or parameters of the items. A commonly used IRT model is the two-parameter logistic model:

$$p_i(\theta) = P(X_i = 1 | \theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \quad (1)$$

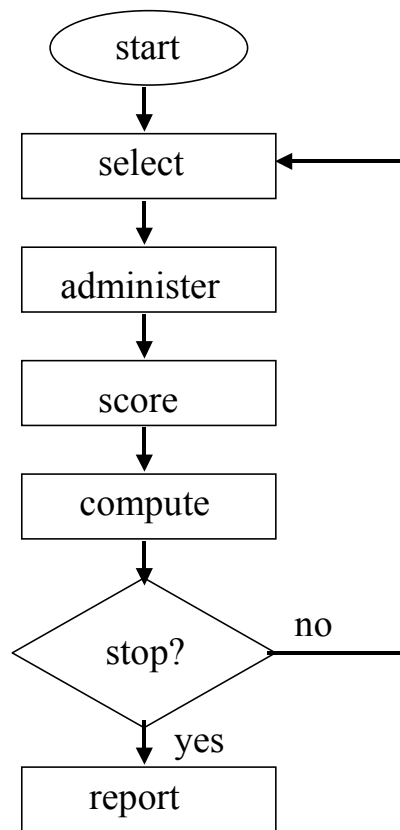
in which  $b_i$  is the location or difficulty parameter and  $a_i$  the discrimination parameter. Because of the favorable item calibration possibilities, this model is in applications often estimated as an OPLM-model (one parameter logistic model), by allowing only for a limited number of fixed discrimination indices  $a_i$  (Verhelst & Glas, 1995). A special case of the two-parameter model is the Rasch model (Rasch, 1960), in which only a difficulty parameter  $b_i$  is present and all  $a_i$  are equal to the same constant.

Working with IRT models starts with item calibration. On the basis of empirical data from administered items an IRT model is fitted and the item parameters and the parameters of one or more ability distributions in the population are estimated. If the item parameters are estimated so accurate as that they can be considered as known, they can be stored in the item bank. The availability of an IRT calibrated item bank is a necessary condition for computerized adaptive testing. The properties of IRT which make it excellently suited for application in CATs are the following.

1. With any subset of the items in a calibrated item bank it is possible to estimate the ability on the same scale. It is therefore not necessary to administer the same items to students in order to get comparable estimates of the ability.
2. The difficulties of the items are expressed on the same scale as the ability of the students. It is therefore possible to adapt a test to a level of a student.
3. The information in an item is a function of the ability, by which the information function can serve as a basis for tailored item selection.

### The testing algorithm

CATs are governed by a testing algorithm. The algorithm is a set of rules which determine the way CATs are started, continued and terminated. In Figure 1.1 a schematic representation of a CAT algorithm is given.



*Figure 1.1. Schematic representation of an adaptive test*

As data on a student's ability are not always available, a CAT starts with presenting a randomly selected item from (a subset of) the item bank to the student. If there is any information on the ability of the student before testing, this information could be used in the first selection. After every administered item an item selection procedure is carried out. From the item bank an item is chosen that is in accordance with the answers given by the student thus far: the test is adapted or tailored to the ability of the tested student. The ability of a student is always inferred on the basis of the likelihood function of the ability  $\theta$ .

Given the scores on  $k$  items,  $X_1 = x_1, X_2 = x_2, \dots, X_k = x_k$ , this function is given by:

$$L(\theta; x_1, x_2, \dots, x_k) = \prod_{i=1}^k L(\theta; x_i) = \prod_{i=1}^k p_i(\theta)^{x_i} (1 - p_i(\theta))^{1-x_i} \quad (2)$$

in which  $p_i(\theta)$  is an IRT model as in (1) in which the item parameters, known from the calibration, are filled in.

There are two main approaches item selection: item selection based on maximum information (Lord, 1970) and Bayesian item selection criteria. Both approaches are reported (Wainer, 2000) to give good results. Bayesian item selection criteria are not used in the studies in this dissertation. The reason for this is that Bayesian criteria in general put more demands on the computer capabilities, which could be a problem in applications. An overview of the Bayesian item selection criteria can be found in Van der Linden (1998).

In maximum information selection approach, that item is selected which has the highest value of the Fisher information function at the current estimate of the ability of the student. The Fisher information function of an item  $i$  is defined by:

$$I_i(\theta) = - \mathcal{E} \left( \frac{\partial^2 \ln L(\theta; x_i)}{\partial \theta^2} \right) \quad (3)$$

If the current ability estimate after administering  $k$  items is  $\hat{\theta}_{(k)}$ , then the next item to be selected from the item bank is the item  $i$  for which

$$\max_i I_i(\hat{\theta}_{(k)}). \quad (4)$$

In the two-parameter model (1), the item information function is given by:

$$I_i(\theta) = a_i^2 p_i(\theta) (1 - p_i(\theta)), \quad (5)$$

which is easily seen to be a single peaked function with the maximum at  $\theta = b_i$ , the difficulty of the item.

The item information function expresses the contribution an item can give to

the accuracy with which the ability of a student is measured. This can be understood by the fact that the item information is additive over items and that the estimate of the standard error of an ability estimate is inversely proportional to the square root of the summed information evaluated at the estimate. The item selection method is largely responsible for the major gain in efficiency of CAT compared to a test with the same items for all students.

After the item has been selected, it is administered to the student and scored. In the subsequent computation phase the student's scores are processed. The likelihood function (2) is used as a basis for making inference on the ability of the student. A frequently used estimator is the maximum likelihood estimator, which is obtained by maximizing (2) with respect to  $\theta$ . In the studies of this thesis the generally statistically superior variant of this estimator, the weighted maximum likelihood estimator (Warm, 1989) is used. This WML estimator is given by:

$$\hat{\theta}_{(k)} = \max_{\theta} \left( \left( \sum_{i=1}^k I_i(\theta) \right)^{\frac{1}{2}} \cdot \prod_{i=1}^k L(\theta; x_i) \right). \quad (6)$$

Besides the ability estimate, an estimation of its standard error is determined. The standard error expresses the accuracy with which the ability of the student is known. This standard error is normally used as a criterion for stopping the testing. If the criterion has not yet been met, a new item is selected, otherwise the result of the test is reported.

In modern CAT applications practical considerations play an important role. That is why, for instance, in the stopping rule of a CAT also specifies a maximum and sometimes a minimum test length. Furthermore several kind of constraints are put on the maximum information selection of items. The main examples of these constraints are with respect to the content of the test and to item exposure. With content constraints a tester wants to employ a desirable content specification for the test, establishing that certain components of the ability that are to be assessed occur in a given proportion (Kingsbury & Zara, 1991). With exposure constraints a twofold problem with unrestricted item

selection can be solved. The first is overexposure: some items are selected so frequently that confidentiality of the items is rapidly and directly compromised. The second is underexposure: there are items in the bank which are so seldomly used that one could wonder how the expense of constructing them can be justified. A recent overview of exposure control methods can be found in Stocking & Lewis (2000).

## **1.2 Overview**

The next chapters in this thesis are a collection of papers which can be read independently of each other. The first three chapters have a more theoretical orientation, and the final three chapters a more practical orientation and are directly applicable in CAT programs.

The first three chapters are devoted to item calibration. Item calibration is critical in CAT, because the item parameters are considered to be known in every step of a CAT algorithm. If the estimates of the item parameters have large standard errors or are possibly changed due to using the items, the validity of inferences made on the basis of a CAT is threatened. A sound item calibration is therefore of utmost importance. In chapter 2, the loss of information in estimating item parameters with conditional maximum likelihood methods, using a general applicable theoretical framework, is treated. In the third chapter, the theory is generalized to incomplete testing designs, which are by necessity applied in case of larger item banks which are used for CATs. The conditions for correct item calibration in different incomplete designs using different item parameter estimation methods are the subject of the fourth chapter.

Traditionally, CAT algorithms are developed for obtaining an efficient estimate of a student's ability. The same algorithms can also be used in classification problems, where not the exact estimate of a student's ability is important, but only the classification in one of a few distinct categories. In this situation a CAT algorithm based on statistical testing, rather than on estimation, can be used. In chapter 5, the use of such algorithms based on the sequential probability ratio test (Wald, 1947) for classification in three distinct categories is studied. In



traditional CATs, item selection methods are based on a criterion which is closely related to statistical estimation. In the chapter 6 an item selection method which is conceptually better related to statistical testing is presented and evaluated.

One of the consequences of optimal item selection in CATs is that it can be expected that each individual will answer about half of the items correct. In the final chapter it is explored whether the selection procedures can be altered in favor of the practical wish to have, for certain groups of examinees, easier or more difficult tests.

### 1.3 References

- Dodd, B.G, De Ayala, R.J. & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19, 5-22.
- Kingsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4, 241-261.
- Lord, F.M. (1970). Some theory for tailored testing. In W.H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. (pp.139-183). New York: Harper and Row.
- Stocking, M.L. & Lewis, C. (2000). Methods for controlling the exposure in CAT. In Van der Linden, W.J. & Glas, C.A.W. (Eds.) (pp. 163-182). *Computerized adaptive testing. Theory and practice*. Dordrecht: Kluwer Academic Publishers.
- Van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63,201-216.
- Van der Linden, W.J. & Hambleton, R.K. (Eds.) (1996). *Handbook of modern item response theory*. New-York: Springer-Verlag.
- Van der Linden, W.J. & Glas, C.A.W. (Eds.) (2000). *Computerized adaptive testing. Theory and practice*. Dordrecht: Kluwer Academic Publishers.
- Van Rijn, P.W., Eggen, T.J.H.M., Hemker, B.T. & Sanders, P.F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 26, 393-411.
- Verhelst, N.D., & Glas, C.A.W. (1995). The one-parameter logistic model. In G.H. Fischer & I.W. Molenaar (Eds.). *Rasch models. Foundations, recent developments, and applications* (pp.215-237). New York: Springer-Verlag.
- Wainer, H. (Ed.) (2000). *Computerized adaptive testing. A primer*. Second edition. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wald, A. (1947). *Sequential analysis*. New York: Wiley.
- Warm, T.A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.



## Chapter 2

### On the loss of information in conditional maximum likelihood estimation of item parameters<sup>1</sup>

---

<sup>1</sup>This chapter was published as Eggen, T.J.H.M.(2000). On the loss of information in conditional maximum likelihood estimation of the item parameters. *Psychometrika*, 65, 337-362., but some major revisions took place.

## **Abstract**

In item response models of the Rasch type (Fischer & Molenaar, 1995), item parameters are often estimated by the conditional maximum likelihood (CML) method. This paper addresses the loss of information in CML estimation by using the information concept of F-information (Liang, 1983). This concept makes it possible to specify the conditions for no loss of information and to define a quantification of information loss. For the dichotomous Rasch model, the derivations will be given in detail to show the use of the F-information concept for making comparisons for different estimation methods. It is shown that by using CML for item parameter estimation, some information is almost always lost. But compared to JML (joint maximum likelihood) as well as to MML (marginal maximum likelihood) the loss is very small. The reported efficiency in the use of information of CML to JML and to MML in several comparisons is always larger than 92%.

## 2.1 Introduction

The Rasch model for measuring a latent trait  $\theta$  with dichotomously scored items, with the responses  $X_{vi} = x_{vi}$  (0 or 1), persons  $v = 1, \dots, n$  and items  $i = 1, \dots, k$  is given by

$$p(x; \omega) = \prod_{v=1}^n \prod_{i=1}^k P(X_{vi} = x_{vi}; \beta_i, \theta_v), \quad (1)$$

$$\text{with } P(X_{vi} = x_{vi}; \beta_i, \theta_v) = \frac{\exp x_{vi}(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}, \quad (2)$$

and  $\omega^T = (\beta^T, \theta^T)$ ,  $\beta^T = (\beta_1, \dots, \beta_k)$  is the vector item parameter (difficulty) and  $\theta^T = (\theta_1, \dots, \theta_n)$  the vector person parameter (ability).

In Fischer and Molenaar (1995) the foundations, the main results and an overview of recent developments and extensions of the Rasch model are given. In these type of item response theory (IRT) models conditional maximum likelihood estimation (CML) is a popular method for estimating item parameters. Although properties as consistency, asymptotic normality and sampling independence of CML estimators are well established, the justification of CML estimation with respect to the possible loss of information due to conditioning is not clear. This topic is addressed in this chapter.

### 2.1.1 Estimation of item parameters in the Rasch model

For the dichotomous Rasch model (1), three likelihood based methods for item parameter estimation are available (Holland, 1990; Molenaar, 1995).

In the first method, joint maximum likelihood (JML), the item parameters are estimated by maximizing (1) with respect to  $\omega$ , given the data  $x$ . It is well known that estimating the item parameters by JML leads to inconsistent estimators (Andersen, 1973). This is caused by the fact that we have a problem in which a limited number of parameters ( $\beta$ ) of interest (items) are to be estimated in the presence of many nuisance (ability) parameters ( $\theta$ ). Eliminating the nuisance parameters gives the solution for this problem (Basu, 1977). In IRT

modeling this elimination is accomplished by the marginal or the conditional maximum likelihood method.

In the marginal maximum likelihood (MML) method it is assumed that the abilities  $\theta_v$  in (1) constitute a random sample from an ability distribution  $h(\theta; \xi)$ , with  $\xi$  the parameters of the ability distribution. In the Rasch model the joint probability of the item responses can then be written as

$$p_m(x; \beta, \xi) = \int p(x | \theta; \beta, \xi) h(\theta; \xi) d\theta = \prod_{v=1}^n \int \prod_{i=1}^k P(X_{vi} = x_{vi} | \theta_v; \beta_i) h(\theta_v; \xi) d\theta_v = L_m(\beta, \xi; x). \quad (3)$$

This so called marginal likelihood function is maximized with respect to  $\beta$  and  $\xi$  in order to get estimates of the item parameters, and, for instance, in the case of a normal ability distribution, two distribution parameters. In the MML method, the nuisance parameters are eliminated by integrating them out. In (3)  $P(X_{vi} = x_{vi} | \theta_v; \beta_i)$  is given by (2).

The MML approach can also be used without specifying a parametric form of the ability distribution, and then estimating this nonparametric distribution together with the item parameters from the data. Results on the estimation in the semiparametric Rasch model can be found in De Leeuw and Verhelst (1986), Lindsay, Clogg and Greco (1991) and Pfanzagl (1993, 1994). The semiparametric models will not be considered explicitly in this chapter.

In Rasch type of IRT models, the CML method is an alternative solution to the inconsistency problem. If there exist sufficient statistics for the nuisance parameters, the model can be separated in a conditional part which is only dependent on the parameters of interest and a part which models the sufficient statistics.

In the dichotomous Rasch model the sum score on the items  $T_v = \sum_{i=1}^k X_{vi}$  is sufficient for  $\theta_v$ ,  $v = 1, \dots, n$ , so we can rewrite (1) as

$$p(x; \omega) = \prod_{v=1}^n f(x_v | t_v; \beta) \cdot \prod_{v=1}^n g(t_v; \beta, \theta_v) \quad (4)$$

with  $X_v = (X_{v1}, \dots, X_{vk})$  the response vector of person  $v = 1, \dots, n$ .

Maximizing, with respect to  $\beta$ , the conditional likelihood

$$L_c(\beta; x | t) = \prod_{v=1}^n f(x_v | t_v; \beta) \quad (5)$$

leads, under mild conditions, to consistent estimators of the item parameters (Andersen, 1973; Pfanzagl, 1994).

### 2.1.2 Information and efficiency

In statistical inference, the Fisher information concept plays an important role. In a distribution  $p(x; \omega)$  depending on one parameter  $\omega$  it is defined as:

$$I_p(\omega) = \mathcal{E}(\partial \ln p(X; \omega) / \partial \omega)^2 \quad (6)$$

It is an intrinsic measure of the accuracy of a distribution (Rao, 1973, p.331); it is the extent by which the uncertainty about the parameter  $\omega$  is reduced by the outcome of a random draw from the distribution. In the evaluation of the quality of estimators the Fisher information has a clear interpretation: it gives a lower bound for the variance of any estimator  $S(X)$  of  $\omega$ :

$$\text{Var } S \geq (1 + \partial b_s(\omega) / \partial \omega)^2 / I_p(\omega),$$

in which  $b_s(\omega) = \mathcal{E}S - \omega$  is the bias of the estimator. If the estimator is unbiased this inequality specializes to the Cramér-Rao inequality:  $\text{Var } S \geq 1 / I_p(\omega)$ .

In the case of several parameters this result is generalized (Rao, 1973): in the class of unbiased estimators the covariance matrix of any estimator is bounded by the inverse of the Fisher information matrix. Estimators reaching the Cramér-Rao lower bound, are called efficient. In likelihood inference, in the evaluation of the estimators the focus is not on unbiasedness, because maximum likelihood estimators (MLE) need not to be unbiased and it is known that if there exists an efficient estimator the MLE procedure will produce it. Instead the large sample



properties of the estimators, the consistency and the asymptotic distribution of the estimators, are evaluated.

It has been shown (Andersen, 1973) that the CML estimators for the item parameters in the Rasch model are under mild conditions consistent and asymptotically normally distributed. If the distribution of the abilities is completely unknown, they are (Pfanzagel, 1994) even asymptotically efficient. The same is true for the MML estimators of the item parameters. In the semiparametric Rasch model, the asymptotic equivalence of the CML and MML item parameter estimators has been proven. (De Leeuw & Verhelst, 1986; Pfanzagel, 1993).

However, these asymptotic results do not imply that there is no loss of information if CML estimation is applied. The point is that by using the conditional likelihood (5) for estimating the item parameters, the second part of the full likelihood (4), which is the marginal distribution of  $T$ , is neglected. And this distribution possibly contains some information on the item parameters. In MML estimation, where no information is discarded, on the other hand, a correct specification of the ability distribution is needed and if this distribution is not the correct one, the resulting estimates of the item parameters can be biased. Furthermore, in MML the loss in information on the item parameter estimation due to the joint estimation of them with the parameters of the ability distribution is not clear.

In psychometrics, there is a general awareness of the possible pre-asymptotic loss of information in estimating the item parameters with CML, however without a quantification of this loss. In case the ability distribution has a known parametric form, Pfanzagel (1994) reports CML to be asymptotically inefficient. And only in the very special case in which it is assumed that the ability distribution parameters are known, Engelen (1989) has shown that the loss in CML estimation compared to MML is small.

In this study a general treatment is given of an information concept, called F-information, which makes it possible to define a clear measure of information loss in using CML estimation. The conditions for no loss of information in

separable models, like the Rasch model, will be given. The theory will be clarified with some examples. In particular, attention will be paid to the dichotomous Rasch model as specified in (1) and (2) and to the Rasch Poisson Counts model for misreadings (Rasch, 1960). For the dichotomous Rasch model the derivations will be given in detail to show the use of the general information concept for making comparisons between different estimation methods. Results will be given of the comparison of CML with JML, and of the comparison of CML with MML in case of a normal ability distribution.

## 2.2 Notation and terminology

Let  $p(x;\omega)$  denote the density (or probability in the discrete case) of a scalar or vector random variable  $X$ , with parameter(vector)  $\omega \in \Omega$ . Generally the parameters of the distribution  $\omega$  can be partitioned in a parameter of interest  $\psi$  and a nuisance parameter  $\tau$ . Both parameters  $\psi$  and  $\tau$  can be vectors. The parameter of interest  $\psi = \psi(\omega)$  has domain  $\Psi$  and the nuisance parameter  $\tau$ , the complementary part to  $\psi$  of  $\omega$ , has domain  $\mathbf{H}$ . It is assumed that  $\Omega = \Psi \times \mathbf{H}$ . Let  $T = T(X)$  be a (vector) statistic,  $f(. | .; .)$  a conditional density or probability of the data given a statistic and  $g(.; .)$  the (marginal) density (or probability) of the statistic. Then in general the probability of the data can be factored as

$$p(x;\omega) = f(x | t;\omega) \cdot g(t;\omega). \quad (7)$$

In case  $T(X)$  is sufficient for  $\tau$  then (7) can be written as

$$p(x;\omega) = f(x | t;\psi) \cdot g(t;\omega). \quad (8)$$

It is a product of a conditional distribution, which only depends on the parameter of interest  $\psi$ , and the distribution of the sufficient statistic for the nuisance parameter  $\tau$  which depends possibly on both  $\psi$  and  $\tau$ . In this case, the model is called *separable* because the first part of the factorization in (8) can be separated for inference on the parameter of interest  $\psi$  from the full model. In CML

estimation the inference is based on only the conditional distribution in (8), while the other part is neglected.

### 2.3 The F-information: definition and basic properties

F-information is a generalization of the Fisher information. In a distribution  $p(x;\omega)$  with an  $m$ -dimensional parameter  $\omega^T = (\omega_1, \dots, \omega_m)^T$  the *Fisher information matrix*, a generalization of (6), is defined as the  $m \times m$  matrix

$$I_p(\omega) = \mathcal{E}[S_{p;\omega} \cdot S_{p;\omega}^T], \quad (9)$$

in which

$$S_{p;\omega} = S_{p;\omega}(X) = \frac{\partial \ln p(X;\omega)}{\partial \omega} \quad (10)$$

is the *efficient score statistic* ( $m \times 1$ -vector) of a distribution  $p$  with respect to  $\omega$ .

In the case that in a distribution  $p(x;\omega)$  the parameters can be partitioned in parameters of interest  $\psi$  and nuisance parameters  $\tau$  the F-information is defined.

#### Definition 1

In a distribution  $p(x;\omega)$ , with  $\omega^T = (\psi^T, \tau^T)$ ,  $\psi^T = (\psi_1, \dots, \psi_k)$  being the parameter of interest and  $\tau^T = (\tau_1, \dots, \tau_n)$  the nuisance parameter, the *F-information* for  $\psi$  in  $p$  is given by

$$I_p(\psi;\omega) = \underset{N}{\text{Min}} \mathcal{E}[m(N) \cdot m(N)^T], \text{ with } m(N) = S_{p;\psi} - N^T \cdot S_{p;\tau}. \quad (11)$$

In (11),  $N$  is a  $n \times k$ -matrix of constants and **Min** is the minimal matrix  $M_1$  in the class **M** of non-negative definite (nnd) matrices which can be written as  $\mathcal{E}[m(N) \cdot m(N)^T]$  for some  $N$ , that is for any  $M \in \mathbf{M} : M - M_1$  is nnd.

F-information is the information in a distribution  $p(x;\omega)$ , which also depends on some nuisance parameters, with respect to only the parameters of interest  $\psi$ . It is a measure of the information for only the interest parameter in which is taken care of the relation between the interest and the nuisance parameter in the

distribution. The concept of F-information, originating from Efron (1977), was first defined by Liang (1983) for scalar parameters and was generalized to vector parameters by Bhapkar (1989).

The F-information concept will now be illuminated by some properties. It should be understood that these properties yield under general regularity conditions, which entail the justification of interchanging the order of differentiation and integration (taking expectations).

1. F-information is a generalization of the Fisher information.

Partition the parameter vector in itself and an empty part  $\omega^T = (\omega^T, \emptyset)$  and observe that  $I_p(\omega; \omega) = \text{Min}_N \mathcal{E}[(S_{p;\omega} - N^T S_{p;\emptyset}) \cdot (S_{p;\omega} - N^T S_{p;\emptyset})^T] = \mathcal{E}[S_{p;\omega} \cdot S_{p;\omega}^T] = I_p(\omega)$ .

2. F-information can also be defined in terms of Fisher information.

It can be shown (Bhapkar, 1989) that if the Fisher information matrix (9) is positive definite and is rewritten as a partitioned matrix:

$$I_p(\omega) = \mathcal{E}[S_{p;\omega} \cdot S_{p;\omega}^T] = \mathcal{E} \begin{bmatrix} S_{p;\psi} S_{p;\psi}^T & S_{p;\psi} S_{p;\tau}^T \\ S_{p;\tau} S_{p;\psi}^T & S_{p;\tau} S_{p;\tau}^T \end{bmatrix} = : \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \quad (12)$$

and its inverse as

$$I_p^{-1}(\omega) = \begin{bmatrix} (I^{-1})_{11} & (I^{-1})_{12} \\ (I^{-1})_{21} & (I^{-1})_{22} \end{bmatrix},$$

then the F-information matrix for  $\psi$  in  $p$  is equal to

$$I_p(\psi; \omega) = I_{11} - I_{12} I_{22}^{-1} I_{21} = ((I^{-1})_{11})^{-1}, \quad (13)$$

which is the inverse of the upper left submatrix of the inverse of the Fisher information matrix.

Suppose both  $\psi$  and  $\tau$  are scalar, then  $N = N^T$  and all efficient scores are scalar. The F-information is then found by minimizing  $\mathcal{E}(S_{p,\psi}^2 - 2N S_{p,\psi} S_{p,\tau} + N^2 S_{p,\tau}^2)$  with respect to  $N$ . After differentiation, the minimum is easily seen to be reached

at  $N = \mathcal{E}(S_{p,\psi} S_{p,\tau}) \cdot (\mathcal{E}(S_{p,\tau})^2)^{-1}$ , and the F-information is given by (13). Generalizing to the multidimensional case, the same result follows for:  $N^T = \mathcal{E}(S_{p,\psi} S_{p,\tau}^T) \cdot [\mathcal{E}(S_{p,\tau} S_{p,\tau}^T)]^{-1}$ .

### 3. F-information has a clear geometrical interpretation.

To enable a geometrical interpretation of the F-information observe:  $I_p(\psi; \omega)$  is the minimal matrix in the class  $\mathbf{M}$  of non-negative definite  $k \times k$  matrices:  $\forall K \in \mathbf{M} : K - I_p(\psi; \omega) \in \mathbf{M}$ . Consider a real valued function  $U$  from  $\mathbf{M} \rightarrow \mathbb{R}$ , so from the class of nnd matrices to the real line. Suppose this function satisfies:

$$\text{a. } U(K) = 0 \text{ if and only if } K = \mathbf{0} \text{ and} \quad (14)$$

$$\text{b. for every } K_1, K_2 \in \mathbf{M} \text{ with } K_1 - K_2 \in \mathbf{M} \text{ implies } U(K_1) \geq U(K_2). \quad (15)$$

Then it follows from (15) that for every  $K \in \mathbf{M}$ , by definition  $I_p(\psi; \omega) \in \mathbf{M}$  and  $K - I_p(\psi; \omega) \in \mathbf{M}$ , which implies  $U(K) \geq U(I_p(\psi; \omega))$ .

So it can be concluded that finding the F-information is equivalent to finding the minimum of the function  $U$ . A matrix function which satisfies (14) and (15) is  $\text{tr}(\cdot)$ , the trace function. So minimizing  $\mathcal{E}[(S_{p,\psi} - N^T S_{p,\tau})(S_{p,\psi} - N^T S_{p,\tau})^T]$ , see (11), is equivalent to minimizing the trace of this matrix and then from standard matrix algebra it follows

$$\text{Min}_N \text{tr}\{ \mathcal{E}[(S_{p,\psi} - N^T S_{p,\tau})(S_{p,\psi} - N^T S_{p,\tau})^T] \} =$$

$$\text{Min}_N \text{tr}\{ \mathcal{E}[(S_{p,\psi} - N^T S_{p,\tau})^T \cdot (S_{p,\psi} - N^T S_{p,\tau})] \} = \text{Min}_N \mathcal{E}[(S_{p,\psi} - N^T S_{p,\tau})^T \cdot (S_{p,\psi} - N^T S_{p,\tau})].$$

The last expression makes it possible to give the F-information the same geometric interpretation as a least squares solution in a regression problem. As can be seen in Figure 2.1, the F-information follows from projecting the efficient score of the parameter of interest onto the space of the efficient score of the nuisance parameter.

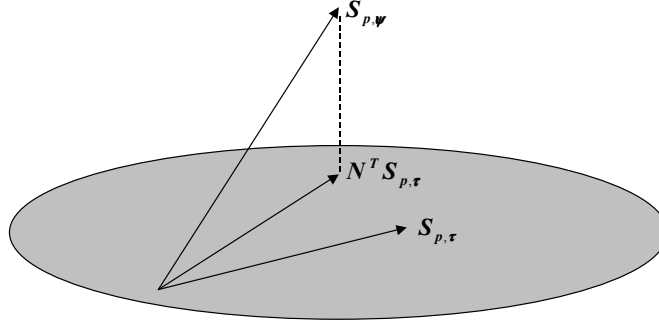


Figure 2.1. Geometrical interpretation of the F-information

#### 4. F-information is the Cramér-Rao bound.

For a finite number of parameters,  $\mathbf{I}_p(\boldsymbol{\psi}; \boldsymbol{\omega})$  is the Cramér-Rao lower bound in the class of unbiased estimators of  $\boldsymbol{\psi}$ , that is, for every covariance matrix  $V$  of unbiased estimators of  $\boldsymbol{\psi}$ ,  $V - (\mathbf{I}_p(\boldsymbol{\psi}; \boldsymbol{\omega}))^{-1}$  is non-negative definite (Rao, 1973, p.326). So when an unbiased estimator of  $\boldsymbol{\psi}$  in a distribution  $p(\mathbf{x}; (\boldsymbol{\psi}, \boldsymbol{\tau}))$  is evaluated with respect to reaching this lower bound, it is seen that all the relevant information on  $\boldsymbol{\psi}$  in this distribution is included in the F-information.

#### 5. Comparing F-information

Like the Fisher information, the F-information as a measure of accuracy can be used to compare the amount of information with respect to the parameter of interest  $\boldsymbol{\psi}$  in different models. The interpretation of this comparison is straightforward in the evaluation of the quality of an estimator of  $\boldsymbol{\psi}$  using these models. This is most convenient of course in case these estimators are unbiased or at least consistent, because then the inverse of the F-information is the lower bound for the covariance matrix of the unbiased estimators. In case the estimators

are biased, this lower bound interpretation of the F-information is limited and only true within the class of estimators with the same bias.

## **2.4 A scalar measure of information**

The F-information for a parameter  $\psi$ , but also the Fisher information, is in general a  $k \times k$ -matrix. And comparing matrices is not straightforward. For the purpose of comparison a scalar measure of information is needed. In this measure in some respect the magnitude of the information should be expressed. Bhapkar (1989, p.147) suggests that a useful scalar measure of information can be defined as a function on the information matrix to the real line satisfying the conditions (14) and (15). In the theory on information functions, being functions from an information matrix to the real line, as developed by Pukelsheim (1993), these two conditions are part of a larger set conditions, which make a function an useful optimality criterion. In Pukelsheim's theory the trace and the determinant of an information matrix meet these conditions. In case the determinant is used the information matrix should be positive definite. Although the trace function of a matrix is very easily computed, there are several good reasons for using the determinant criterion for comparing the information matrices. In chapter 3 of this dissertation it is shown that for comparing information matrices for different estimation methods in the Rasch model the determinant has some unique properties. Here only a general motivation for using the determinant criterion will be given.

In multivariate linear model theory the determinant of the variance covariance matrix  $V$  which equals the inverse of the information matrix,  $I = V^{-1}$ , has a well established meaning and is called the generalized variance. This is because under normality assumptions in these models the confidence ellipsoid for an estimatable contrast of the parameters, using an optimal estimator, has a volume which is inversely proportional to  $(\det I)^{1/2}$ . Hence a large value of  $\det I$  ensures a small volume of the confidence ellipsoid and maximizing the determinant of the information matrix is the same as minimizing the generalized variance, because  $(\det I)^{-1} = \det(I^{-1})$ . This implies that a larger determinant of the information

matrix, the less the uncertainty is about the (contrast of the) parameters of the model.

In optimality theory as a rule not the determinant of a  $k \times k$  information matrix is considered but  $(1/k)^{\text{th}}$  power of the determinant. Although both induce the same ordering of the information matrices, the standardized measure is independent of the dimension of the model and has some theoretical advantages (Pukelsheim, 1993, p.136). For this reason in this study also  $(\det I)^{1/k}$  will be used as a scalar measure of information.

For comparing the (F-)information in two models the ratio of the scalar measure of information is in common use and is defined as the *information efficiency*. Having a  $k$ -vector parameter of interest  $\psi$ , the F-informations in two models  $f$  and  $p$  are  $k \times k$  matrices  $I_f(\psi; \omega)$  and  $I_p(\psi; \omega)$  then the information efficiency of two models with respect to  $\psi$  will be computed as:

$$\text{INFEFF}(\psi; f:p) = \left( \frac{\det(I_f(\psi; \omega))}{\det(I_p(\psi; \omega))} \right)^{1/k} \quad (16)$$

In applications the information matrices are not always of full rank and a proper normalization is to be chosen in order to compute the determinants. It will be clear that in comparing information matrices of different models in both the same normalization has to be used.

## 2.5 F-information in separable models

The F-information can be used to quantify the loss of information in estimating the parameters of interest  $\psi$  by just considering the conditional distribution  $f(x|t; \psi)$ , in case the original model can be factored as in (8) as  $p(x; \omega) = f(x | t; \psi) \cdot g(t; \omega)$ . Moreover, conditions can be specified for no loss of information using this method. In these separable models some theorems on the F-information will be shown to hold. Throughout it is assumed that regularity conditions are met which guarantee the existence of the Fisher information and allow interchanging the order of differentiation and integration of the logarithm of the model. In order to be able to prove these theorems on the F-information,



some properties of the efficient score statistics, defined in (10) and involved in the definition of F-information, are needed and will be given first.

### 2.5.1 Properties of the efficient score statistics

Three basic properties of the efficient score statistics are:

*Property 1*

$$\mathcal{E} S_{p;\Psi} = \mathcal{E} S_{p;\tau} = 0.$$

*Property 2*

$$\text{a. } \text{Cov}(S_{p;\Psi}, S_{p;\Psi}^T) = \mathcal{E} (S_{p;\Psi} S_{p;\Psi}^T) = -\mathcal{E} S_{p;\Psi, \Psi^T} := -\mathcal{E} \left( \frac{\partial^2 \ln p}{\partial \Psi \partial \Psi^T} \right); \text{ and}$$

$$\text{b. } \text{Cov}(S_{p;\tau}, S_{p;\tau}^T) = \mathcal{E} (S_{p;\tau} S_{p;\tau}^T) = -\mathcal{E} S_{p;\tau, \tau^T}.$$

*Property 3*

$$\text{a. } \text{Cov}(S_{p;\Psi}, S_{p;\tau}^T) = \mathcal{E} (S_{p;\Psi} S_{p;\tau}^T) = -\mathcal{E} S_{p;\Psi, \tau^T} := -\mathcal{E} \left( \frac{\partial^2 \ln p}{\partial \Psi \partial \tau^T} \right); \text{ and}$$

$$\text{b. } \text{Cov}(S_{p;\tau}, S_{p;\Psi}^T) = \mathcal{E} (S_{p;\tau} S_{p;\Psi}^T) = -\mathcal{E} S_{p;\tau, \Psi^T}.$$

These three properties are valid in any distribution with parameter  $\omega^T = (\Psi^T, \tau^T)$ . So when a decomposition as in (7) or (8) is considered, the properties also apply to the efficient score statistics of  $f(x | t; \omega)$  and  $g(t; \omega)$ . Then the following properties, relating the efficient score statistics of  $p$ ,  $f$  and  $g$ , are easily deduced. The proofs are in the appendix of this chapter (p.59).

*Property 4*

$$\text{a. } \mathcal{E} (S_{p;\Psi} | T = t) = S_{g;\Psi}; \quad \text{and b. } \mathcal{E} (S_{p;\tau} | T = t) = S_{g;\tau}.$$

*Property 5*

$$\text{a. } \mathcal{E} (S_{p;\Psi} S_{g;\Psi}^T) = \mathcal{E} (S_{g;\Psi} S_{g;\Psi}^T); \quad \text{and b. } \mathcal{E} (S_{p;\tau} S_{g;\tau}^T) = \mathcal{E} (S_{g;\tau} S_{g;\tau}^T).$$

In the case of a separable distribution that can be decomposed as in (8), the following properties on the relation between the score statistics with the respect to the same parameters can be added. Property 6a. is true by definition, while 6b. is a special case of the definition.

*Property 6*

$$\text{a. } S_{p;\psi} = S_{f;\psi} + S_{g;\psi}; \quad \text{and b. } S_{f;\tau} = 0 \text{ and } S_{p;\tau} = S_{g;\tau}.$$

By application of property 6, and using the properties 3 and 5, the covariances of the efficient score statistics of  $f$  and  $g$  with respect to different parameters are shown to be 0.

*Property 7*

$$\text{a. } \mathcal{E}(S_{f;\psi} S_{g;\tau}^T) = 0; \quad \text{and b. } \mathcal{E}(S_{f;\tau} S_{g;\psi}^T) = 0.$$

The same is true for the covariances of the efficient score statistics of  $f$  and  $g$  with respect to the same parameters.

*Property 8*

$$\text{a. } \mathcal{E}(S_{f;\psi} S_{g;\psi}^T) = 0; \quad \text{and b. } \mathcal{E}(S_{f;\tau} S_{g;\tau}^T) = 0$$

### ***2.5.2 Theorems on the F-information in separable models***

In the first theorem, the additivity of the F-informations and the relation between the F-information in the conditional distribution and the conditional Fisher information is given.

### Theorem 1

For a separable distribution which factors in  $p(x; \omega) = f(x | t; \psi) \cdot g(t; \omega)$

- a. the F-information for  $\psi$  in  $p$  is the sum of the F-information in  $f$  and the F-information in  $g$ :

$$I_p(\psi; \omega) = I_f(\psi; \omega) + I_g(\psi; \omega). \quad (17a)$$

- b. Moreover, the F-information in the conditional distribution is the same as the expectation of the Fisher information in the conditional distribution, with the expectation taken with respect to the distribution of the sufficient statistic  $T$ :

$$\begin{aligned} I_f(\psi; \omega) &= \mathcal{E} I_f(\psi | T). \text{ So,} \\ I_p(\psi; \omega) &= \mathcal{E} I_f(\psi | T) + I_g(\psi; \omega). \end{aligned} \quad (17b)$$

The proof of the theorem is in the appendix of this chapter (p.60).

In separable distributions therefore, the F-information with respect to the parameter of interest  $\psi$  can be written as the sum of the F-information in the conditional distribution and the F-information in the marginal distribution of the statistic  $T(X)$ . This additivity of the components of the F-information is very attractive. It will be clear that the loss of information by using the conditional model instead of the full model in an inference on  $\psi$  is given by the F-information in  $T$ :  $I_g(\psi; \omega)$ , the F-information in the neglected part of the full model. The information efficiency defined in (16) as the ratio of the  $(1/k)^{\text{th}}$  power of the determinants of the F-information in the conditional model and in the full model, can also be computed.

### Example

An example of the computation of the F-information is given for a problem in which both the interest and the nuisance parameter are scalar. Consider a random sample from a normal distribution  $X = (X_1, \dots, X_n)$ ,  $X_i \sim N(\mu, \sigma^2)$ , with both the mean  $\mu$  and the variance  $\sigma^2$  unknown. The interest is in estimating  $\sigma^2$ , while  $\mu$

is considered as a nuisance parameter. In this problem, the sample mean  $T = \sum_i X_i / n$  is sufficient for  $\mu$  and the full model can be decomposed as in (8):

$$p(x; \mu, \sigma^2) = f(x | t; \sigma^2) g(t; \mu, \sigma^2).$$

For estimating  $\sigma^2$ , the full model  $p$  or the conditional model  $f$  can be used.

The Fisher information matrices for the full model  $p$  and the distribution  $g$  of  $T$ , ( $T \sim N(\mu, \sigma^2/n)$ ), are

$$I_p(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix} \text{ and } I_g(\mu, \sigma^2) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}.$$

The F-information with respect to  $\sigma^2$  in  $p$  and  $g$  follows from (13):

$I_p(\sigma^2; (\mu, \sigma^2)) = n/2\sigma^4$ , and  $I_g(\sigma^2; (\mu, \sigma^2)) = 1/2\sigma^4$ . From Theorem 1 (17a), the F-information in the conditional model equals:  $I_f(\sigma^2; (\mu, \sigma^2)) = (n-1)/2\sigma^4$ . So in using CML instead of ML for estimating  $\sigma^2$ , there is a small loss of information which amounts to  $1/2\sigma^4$ . The information efficiency of CML versus ML is:  $I_f(\sigma^2; (\mu, \sigma^2)) / I_p(\sigma^2; (\mu, \sigma^2)) = (n-1)/n$ . In this example it can be remarked that, although there is a little loss of information in using CML, the resulting CML estimator is unbiased, whereas the ML estimator is not:  $\hat{\sigma}_{cml}^2 = \sum_i (x_i - t)^2 / (n-1)$  and  $\hat{\sigma}_{ml}^2 = \sum_i (x_i - t)^2 / n$ .  $\square$

#### The conditions for no loss of information

Next, two definitions are given which make it possible to specify the conditions under which there is no loss of information when the conditional model is used instead of the full model.

*Definition 2*

In a separable distribution, satisfying (8), the statistic  $T = T(X)$  is *strongly ancillary* for  $\Psi$ , if  $g(t; \omega)$  only depends on  $\tau$ :  $p(x; \omega) = f(x | t; \Psi) \cdot g(t; \tau)$ .

It is easily understood that if the statistic  $T$  is strongly ancillary for  $\Psi$ , there will be no loss of information if in an inference on  $\Psi$  the conditional distribution  $f(x | t; \Psi)$  is used instead of  $p(x; \omega)$ . In this case the part which is neglected does not depend on the parameter  $\Psi$  and  $I_g(\Psi; \omega) = 0$ .

Bhaskar (1989), generalizing an earlier result of Liang (1983) for scalar parameters to vector parameters, has proven that there is also no loss of information if weaker conditions on the distribution of  $T$  are fulfilled. Consider the following definition:

*Definition 3*

In a separable distribution, satisfying (8), the statistic  $T = T(X)$  is *weakly ancillary* for  $\Psi$ , if there exists a one-to-one reparametrization between  $(\Psi, \tau)$  and  $(\Psi, \delta)$  such that  $g(t; \omega) = g(t; \delta)$ , that is, only depends on  $\delta$ .

*Example*

$X_1$  and  $X_2$  are independent Poisson distributed variables.  $X = (X_1, X_2)$ ,  $\mathcal{E}(X_1) = \lambda$  and  $\mathcal{E}(X_2) = \lambda\rho$ . The interest and nuisance parameter are respectively  $\ln \rho$  and  $\ln \lambda$ .  $T(X) = X_1 + X_2$  is also Poisson distributed with  $\mathcal{E}(T) = \lambda + \lambda\rho$ . Factorization (8) yields:

$$p(x; \lambda, \rho) = \frac{\exp[(x_1 + x_2) \ln \lambda + x_2 \ln \rho - (\lambda + \lambda\rho)]}{x_1! x_2!} =$$

$$\frac{t! \exp[x_2 \ln \rho]}{x_1! x_2! \exp[t \ln(1 + \rho)]} \cdot \frac{\exp[t \ln \lambda + t \ln(1 + \rho) - (\lambda + \lambda\rho)]}{t!} = f(x | t; \rho) \cdot g(t; \lambda, \rho)$$

Apply the one-one reparametrization  $\ln \rho = \ln \rho$  and  $\ln \lambda = \ln \delta - \ln(1 + \rho)$ . The distribution of  $T$  is then given by:  $g(t; \rho, \delta) = (\exp[t \ln \delta - \delta]) / t!$ , which only depends on  $\delta$ .  $\square$

Next Bhapkar's theorem (1989), specifying the sufficient conditions for no loss of information, is given.

### *Theorem 2*

For a separable distribution which factors in  $p(x; \omega) = f(x | t; \psi) \cdot g(t; \omega)$  and in which the statistic  $T = T(X)$  is weakly ancillary for  $\psi$ , there is no loss of information in using the conditional distribution  $f(x | t; \psi)$  instead of  $p(x; \omega)$  for inference on  $\psi$ :  $I_g(\psi; \omega) = 0$ .

The condition of weak ancillarity, although slightly differently defined, was also the key condition under which Andersen (1970, 1973) established the asymptotic efficiency of conditional maximum likelihood estimators. According to Andersen (1973, p 99), weak ancillarity of the statistic  $T$  means that no inference about the parameter of interest  $\psi$  can be drawn from the distribution of  $T$ ,  $g(t; \omega)$ , that is not completely dependent on the specification of the nuisance parameter  $\tau$ . In other words, nothing can be learned from the data about the parameter of interest from the sole observation of the statistic. Although intuitively weak ancillarity may be an appealing concept, it is not easy to show in general that a statistic has this property. However, for models belonging to the exponential family, necessary and sufficient conditions for weak ancillarity, which are easily checked, are given in the next theorem (Andersen, 1973; Bhapkar 1989; Liang, 1983).

### Theorem 3

If  $X$  has a distribution belonging to the exponential family with natural parameters  $\omega = (\psi, \tau)$ :

$$p(x; \omega) = c(\psi, \tau) \cdot \exp \{u^T(x) \cdot \psi + t^T(x) \cdot \tau + b(x)\} \quad (18)$$

and the distribution of  $T = T(X)$  is given by

$$g(t; \omega) = c(\psi, \tau) \cdot \gamma(t, \psi) \cdot \exp(t^T \tau), \quad (19)$$

then  $T$  is weakly ancillary for  $\psi$  if and only if there exist functions of  $\psi$  only and independent of the data:  $w_i(\psi)$ ,  $i = 1, \dots, k$ , and  $v(\psi)$ , such that:

$$\frac{\partial \ln \gamma(t, \psi)}{\partial \psi_i} = w_i(\psi) t + v(\psi), \text{ for } i = 1, \dots, k \quad (20)$$

for all (ae)  $t$ .

## 2.6 Checking the conditions for no loss of information using CML in two Rasch models

### 2.6.1 The Rasch Poisson counts model

In this model, proposed by Rasch (1960), the number of failures  $X_{vi}$  of person  $v = 1, \dots, n$  on test  $i = 1, \dots, k$ , which each consist of a number of items with low error probabilities, is considered. A well known application of the model is the number of misreadings in texts. The model assumes that  $X_{vi}$  is Poisson distributed with parameter  $\lambda_{vi} = \beta_i \theta_v$ ,  $\beta_i$  being the difficulty parameter of text  $i$  and  $1/\theta_v$  the ability parameter of person  $v$ . Writing the model in the exponential family form (18) gives:

$$P(X_{vi}=x_{vi}; \beta_i \theta_v) = \frac{\exp(-\beta_i \theta_v) \cdot \exp\{x_{vi} \ln \beta_i + x_{vi} \ln \theta_v\}}{x_{vi}!}, \quad x_{vi} = 0, 1, 2, \dots$$

With the assumption of independence over texts and persons we get:

$$p(x; \beta, \theta) = \prod_i \prod_v P(X_{vi} = x_{vi}; \beta_i, \theta_v) =$$

$$\frac{\exp\{\sum_i \sum_v -\beta_i \theta_v\} \cdot \exp\{\sum_i (\sum_v x_{vi}) \ln \beta_i + \sum_v (\sum_i x_{vi}) \ln \theta_v\}}{\prod_i \prod_v x_{vi}!}.$$

It is easily seen that the model is a member of the exponential family (18), in which the statistic  $T_v = \sum_i X_{vi}$ , the number of failures of person  $v$  is sufficient for  $\ln \theta_v$ , for  $v = 1, \dots, n$ .

So, the model is separable:

$$p(x; \beta, \theta) = \prod_{v=1}^n f(x_v | t_v; \beta) \cdot \prod_{v=1}^n g(t_v; \beta, \theta_v),$$

and for estimating the text parameters  $\beta$  CML estimation can be considered to use. Instead of the full model  $p(x; \beta, \theta)$ , only the conditional model, the first part of the factorization, is used. Whether neglecting the distribution of  $T$  causes loss of information is now easily checked.

Observe that  $T_v$  is also Poisson distributed, with parameter  $\theta_v \sum_i \beta_i$ :

$$g(t; \beta, \theta) = \prod_{v=1}^n g(t_v; \beta, \theta_v) =$$

$$\frac{\exp\{\sum_i \sum_v -\beta_i \theta_v\} \cdot \exp\{(\sum_v t_v) \ln \sum_i \beta_i + \sum_v t_v \ln \theta_v\}}{\prod_v t_v!}.$$

So the function  $\gamma(t, \psi)$  in (19) is given by

$$\gamma(t; \beta) = \prod_{v=1}^n (\exp\{t_v \ln \sum_i \beta_i\}) / t_v!,$$

and because  $\ln \gamma(t; \beta) = \sum_v \{t_v \ln \sum_i \beta_i - \ln(t_v!)\}$



$$\frac{\partial \ln \gamma(t; \beta)}{\partial \beta_j} = \sum_v t_v \frac{1}{\sum_i \beta_i}, \text{ for } j = 1, \dots, k$$

The condition, (20), for  $T$  being weakly ancillary for  $\beta$  is seen to be fulfilled:  $\partial \ln \gamma(t; \beta) / \partial \beta_j$  is a linear function of  $t$ , and  $w_j(\psi) = 1 / \sum_i \beta_i$ ,  $j = 1, \dots, k$  only depends on the text parameter  $\beta$ . So, there will be no loss of information if the text parameters are estimated with CML. In terms of F-information, with  $\omega^T = (\beta^T, \theta^T)$ :  $I_p(\beta; \omega) = I_f(\beta; \omega)$  or  $I_g(\beta; \omega) = \mathbf{0}$ .

### 2.6.2 The Rasch model for dichotomously scored items

The same conditions will be checked for the model which was presented in the introduction. Writing this model, (1) and (2), as

$$p(x; \beta, \theta) = \frac{\exp\{-\sum_i (\sum_v x_{vi}) \beta_i + \sum_v (\sum_i x_{vi}) \theta_v\}}{\prod_i \prod_v \{1 + \exp(\theta_v - \beta_i)\}}, \quad (21)$$

is seen to belong to the exponential family.  $T_v = \sum_i X_{vi}$  is sufficient for  $\theta_v$ , for  $v = 1, \dots, n$ . The distribution of  $T$  is checked for weak ancillarity. This distribution is given by

$$g(t; \beta, \theta) = \prod_{v=1}^n g(t_v; \beta, \theta_v) = \prod_{v=1}^n \frac{\exp(\theta_v t_v) \cdot \gamma_{t_v}(\beta)}{\prod_{i=1}^k \{1 + \exp(\theta_v - \beta_i)\}}, \quad (22)$$

$$\text{with } \gamma_{t_v}(\beta) = \sum_{\sum_i x_{vi} = t_v} \exp(-\sum_{i=1}^k \beta_i x_{vi}).$$

For,  $j = 1, \dots, k$ , this gives

$$\frac{\partial \ln \prod_{v=1}^n \gamma_{t_v}(\beta)}{\partial \beta_j} = \sum_{v=1}^n \frac{\partial \ln \gamma_{t_v}(\beta)}{\partial \beta_j} = \sum_{v=1}^n \frac{e^{-\beta_j} \cdot \gamma_{t_v-1}^{(j)}}{\gamma_{t_v}(\beta)} \cdot \frac{t_v}{t_v} = \sum_{v=1}^n \frac{e^{-\beta_j} \cdot \gamma_{t_v-1}^{(j)}}{\sum_{i=1}^k e^{-\beta_i} \cdot \gamma_{t_v-1}^{(i)}} \cdot t_v, \quad (23)$$

in which  $\gamma_{t_v-1}^{(j)} := \partial \gamma_{t_v}(\beta) / \partial e^{-\beta_j}$ , for  $j = 1, \dots, k$ .

Note that the last equality in (23) uses an expression given by Fischer (1974, p.242):  $\gamma_{t_v}(\beta) \cdot t_v = \sum_{i=1}^k e^{-\beta_i} \gamma_{t_v-1}^{(i)}$ .

It is seen that condition (20) of Theorem 3 is not fulfilled, since the function  $w_j(\psi)$  is in general not only dependent on the item parameters  $\beta$ , but also on the statistic  $t_v$ .  $T$  is therefore not weakly ancillary for  $\beta$  the Rasch model. This means that using CML estimation is possibly accompanied with loss of information compared to the situation in which the full model is used. The amount of loss will be explored in the next sections of this paper. It can be noted that there is a special case, in which  $w_j(\psi)$  is independent of the data and there is no loss in CML estimation. In chapter 3 of this dissertation this case is discussed.

## 2.7 F-information in the dichotomous Rasch model: comparing JML and CML

In order to determine the loss of information in the Rasch model, with both the item- and ability parameters considered fixed, the expressions for the F-information in the full model  $p$ , the distribution  $g$  of the sufficient statistic and conditional distribution  $f$ , are given first. The Fisher information matrix of the model  $p(x; \beta, \theta)$ , given in (21), with respect to both parameters  $\beta$  and  $\theta$ , is written in the partitioned form as in (12):

$$I_p(\omega) = \mathcal{E} \begin{bmatrix} S_{p;\beta} S_{p;\beta}^T & S_{p;\beta} S_{p;\theta}^T \\ S_{p;\theta} S_{p;\beta}^T & S_{p;\theta} S_{p;\theta}^T \end{bmatrix} = : \begin{bmatrix} I_p^{\beta\beta^T} & I_p^{\beta\theta^T} \\ I_p^{\theta\beta^T} & I_p^{\theta\theta^T} \end{bmatrix} \quad (24)$$

Using the properties 1, 2 and 3 of the efficient score statistics, the submatrices,  $I_p^{\beta\beta^T} (k \times k)$ ,  $I_p^{\beta\theta^T} = [I_p^{\theta\beta^T}]^T (k \times n)$  and  $I_p^{\theta\theta^T} (n \times n)$ , which determine the F-information (see (13)), are easily seen to be the negative expectation of the second derivatives of  $\ln p$  with respect to the parameters. Define

$$p_{vi} := \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}. \quad (25)$$

Then the elements of the submatrices in (24) are given by:

$$(\mathbf{I}_p^{\beta\beta^T})_{ii} = \sum_{v=1}^n p_{vi}(1 - p_{vi}) \text{ for } i = 1, \dots, k; \quad (26a)$$

$$(\mathbf{I}_p^{\beta\beta^T})_{ij} = 0 \text{ for } i \neq j = 1, \dots, k; \quad (26b)$$

$$(\mathbf{I}_p^{\beta\beta^T})_{iv} = -p_{vi}(1 - p_{vi}) \text{ for } i = 1, \dots, k; v = 1, \dots, n; \quad (26c)$$

$$(\mathbf{I}_p^{\theta\theta^T})_{vv} = \sum_{i=1}^k p_{vi}(1 - p_{vi}) \text{ for } v = 1, \dots, n; \quad (26d)$$

$$(\mathbf{I}_p^{\theta\theta^T})_{vw} = 0 \text{ for } v \neq w = 1, \dots, n. \quad (26e)$$

The Fisher information matrix is clearly singular. For every row and column

$$\sum_{i=1}^{k+n} (\mathbf{I}_p(\omega))_{ij} = \sum_{j=1}^{k+n} (\mathbf{I}_p(\omega))_{ij} = 0 \text{ for } i, j = 1, \dots, k+n.$$

But because the diagonal submatrix  $\mathbf{I}_p^{\theta\theta^T}$  is positive definite, and the elements of the inverse are given by the reciprocals of (26d), using (13), the F-information in the full Rasch model  $\mathbf{I}_p(\beta; \omega)$ , is given by

$$(\mathbf{I}_p(\beta; \omega))_{ii} = \sum_{v=1}^n p_{vi}(1 - p_{vi}) - \sum_{v=1}^n \frac{[p_{vi}(1 - p_{vi})]^2}{\sum_{\ell=1}^k p_{v\ell}(1 - p_{v\ell})} \text{ for } j = 1, \dots, k, \quad (27a)$$

$$(\mathbf{I}_p(\beta; \omega))_{ij} = - \sum_{v=1}^n \frac{p_{vi}(1 - p_{vi}) \cdot p_{vj}(1 - p_{vj})}{\sum_{\ell=1}^k p_{v\ell}(1 - p_{v\ell})} \text{ for } i \neq j = 1, \dots, k. \quad (27b)$$

Next, the F-information in the distribution  $g(t; \beta, \theta)$  of the sufficient statistic (see (22)) is determined. The Fisher information matrix, analogous to (24), is given by

$$I_g(\omega) = \mathcal{E} \begin{bmatrix} S_{g;\beta} S_{g;\beta}^T & S_{g;\beta} S_{g;\theta}^T \\ S_{g;\theta} S_{g;\beta}^T & S_{g;\theta} S_{g;\theta}^T \end{bmatrix} = : \begin{bmatrix} I_g^{\beta\beta^T} & I_g^{\beta\theta^T} \\ I_g^{\theta\beta^T} & I_g^{\theta\theta^T} \end{bmatrix}. \quad (28)$$

Because of the properties 3 and 6 ( $S_{g,\theta} = S_{p,\theta}$ ) of the efficient score statistics,  $I_g^{\beta\theta^T} = I_p^{\beta\theta^T}$  and  $I_g^{\theta\theta^T} = I_p^{\theta\theta^T}$ , which are given by (26c), (26d) and (26e). This is, of course, a consequence of the fact that  $T_v$  is sufficient for  $\theta_v$  in  $p(x; \beta, \theta)$ . The second derivatives of  $\ln g$  with respect to  $\beta$  are given by:

$$\partial^2 \ln g / \partial \beta_i^2 = \sum_{v=1}^n \left\{ \frac{e^{-\beta_i} \cdot \gamma_{t_v-1}^{(i)}}{\gamma_{t_v}} - \left[ \frac{e^{-\beta_i} \cdot \gamma_{t_v-1}^{(i)}}{\gamma_{t_v}} \right]^2 - p_{vi}(1-p_{vi}) \right\} \text{ for } i = 1, \dots, k,$$

and

$$\frac{\partial^2 \ln g}{\partial \beta_j \partial \beta_i} = \sum_{v=1}^n \left\{ \frac{e^{-\beta_i - \beta_j} \cdot \gamma_{t_v-2}^{(ij)}}{\gamma_{t_v}} - \frac{e^{-\beta_i - \beta_j} \cdot \gamma_{t_v-1}^{(i)} \gamma_{t_v-1}^{(j)}}{\gamma_{t_v}^2} \right\} \text{ for } i \neq j = 1, \dots, k,$$

with

$$\gamma_{t_v-2}^{(ij)} := \frac{\partial^2 \gamma_{t_v}(\beta)}{\partial e^{-\beta_i} \partial e^{-\beta_j}}.$$

Observe that in the dichotomous Rasch model

$$P(X_{vi} = 1 \mid T_v = t_v; \beta) = \frac{e^{-\beta_i} \cdot \gamma_{t_v-1}^{(i)}}{\gamma_{t_v}} = : p_{vi|t_v} \text{ for } i = 1, \dots, k \text{ and}$$

$$P(X_{vi} = 1, X_{vj} = 1 \mid T_v = t_v; \beta) = \frac{e^{-\beta_i - \beta_j} \cdot \gamma_{t_v-2}^{(ij)}}{\gamma_{t_v}} = : p_{vi,vj|t_v} \text{ for } i \neq j = 1, \dots, k.$$

So, the elements of the submatrix  $I_g^{\beta\beta^T}$  of (28) are given by:

$$(\mathbf{I}_g^{\beta\beta^T})_{ii} = \mathcal{E} \sum_{v=1}^n [-p_{vi|T_v}(1-p_{vi|T_v}) + p_{vi}(1-p_{vi})], i = 1, \dots, k \text{ and} \quad (29a)$$

$$(\mathbf{I}_g^{\beta\beta^T})_{ij} = \mathcal{E} \sum_{v=1}^n [-p_{vi,vj|T_v} + p_{vi|T_v} \cdot p_{vj|T_v}], i \neq j = 1, \dots, k. \quad (29b)$$

The Fisher information matrix of  $\mathbf{g}$  with respect to  $\boldsymbol{\beta}$  and  $\boldsymbol{\theta}$ , is specified in (29ab) and (26cde). Using (13), the F-information in  $\mathbf{g}$  with respect to  $\boldsymbol{\beta}$ , for  $i = 1, \dots, k$  is then given by:

$$(\mathbf{I}_g(\boldsymbol{\beta}; \boldsymbol{\omega}))_{ii} = \mathcal{E} \sum_{v=1}^n [-p_{vi|T_v}(1-p_{vi|T_v}) + p_{vi}(1-p_{vi})] - \sum_{v=1}^n \frac{[p_{vi}(1-p_{vi})]^2}{\sum_{\ell=1}^k p_{v\ell}(1-p_{v\ell})} \quad (30a)$$

and for  $i \neq j = 1, \dots, k$

$$(\mathbf{I}_g(\boldsymbol{\beta}; \boldsymbol{\omega}))_{ij} = \mathcal{E} \sum_{v=1}^n [-p_{vi,vj|T_v} + p_{vi|T_v} \cdot p_{vj|T_v}] - \sum_{v=1}^n \frac{p_{vi}(1-p_{vi}) \cdot p_{vj}(1-p_{vj})}{\sum_{\ell=1}^k p_{v\ell}(1-p_{v\ell})}. \quad (30b)$$

Using (16ab) on the additivity of the F-information, the F-information in the conditional distribution  $f(\mathbf{x} | \mathbf{t}; \boldsymbol{\beta})$  simply is found by subtracting the expressions in (30ab) from (27ab) and this gives:

$$(\mathcal{E} \mathbf{I}_f(\boldsymbol{\beta} | T))_{ii} = \mathcal{E} \sum_{v=1}^n p_{vi|T_v}(1-p_{vi|T_v}) \text{ for } i = 1, \dots, k \text{ and} \quad (31a)$$

$$(\mathcal{E} \mathbf{I}_f(\boldsymbol{\beta} | T))_{ij} = \mathcal{E} \sum_{v=1}^n (p_{vi,vj|T_v} - p_{vi|T_v} \cdot p_{vj|T_v}) \text{ for } i \neq j = 1, \dots, k. \quad (31b)$$

All three relevant F-information matrices in the Rasch model are now specified in (27ab), (30ab) and (31ab). Once the expectations in (30ab) and (31ab) over the distribution of  $T_v$  (22) are determined, they are easily computed for a given set

of item- and ability parameters. The diagonal of the F-information in the conditional distribution (31a), for instance, is

$$(\mathcal{E} I_f(\beta | T))_{ii} = \sum_{v=1}^n \sum_{t_v=0}^k p_{vi|t_v} (1 - p_{vi|t_v}) \exp(\theta_v t_v) \gamma_{t_v} \prod_{i=1}^k (1 - p_{vi}) \text{ for } i = 1, \dots, k$$

The expressions (30ab) and (31b) change similarly.

Note that the expressions for the F-information matrices are all sums of independent contributions of the persons  $v = 1, \dots, n$ .

### Example

Consider the Rasch model with 3 items,  $\beta_1 = -.5$ ,  $\beta_2 = 0.0$  and  $\beta_3 = 1.75$ , and 4 persons with  $\theta_v$  respectively -1.0, 0.0, 1.0, and 2.0.

The Fisher information matrix ( $k+n \times k+n$ ) is given by:

$$I_p(\omega) = \begin{pmatrix} .689 & .000 & .000 & -.235 & -.235 & -.149 & -.070 \\ .000 & .748 & .000 & -.197 & -.250 & -.197 & -.105 \\ .000 & .000 & .647 & -.056 & -.126 & -.218 & -.246 \\ -.235 & -.197 & -.056 & .488 & .000 & .000 & .000 \\ -.235 & -.250 & -.126 & .000 & .611 & .000 & .000 \\ -.149 & -.197 & -.218 & .000 & .000 & .564 & .000 \\ -.070 & -.105 & -.246 & .000 & .000 & .000 & .421 \end{pmatrix}$$

The structure of the matrix is clear: The item parameter submatrix as well as the ability parameter submatrix are diagonal and all covariances between score statistics with respect to item parameters and ability parameters are negative.

The symmetric F-information matrix in  $p$  ( $k \times k$ ), using (27ab), is given by

$$I_p(\beta; \omega) = \begin{pmatrix} .435 & -.260 & -.174 \\ -.260 & .472 & -.212 \\ -.174 & -.212 & .386 \end{pmatrix}$$

Note that in JML estimation in the Rasch model the diagonal of the item parameter submatrix of the Fisher information matrix is sometimes used for estimating the standard errors of the item parameters (Wright, 1977). (In the computations the matrix is evaluated at the estimates). This practice is in fact based on the assumption that the ability parameters are known and leads to an underestimate of the standard errors. A better estimate, taking account of the negative covariances between the item and ability score statistics, would be to use the diagonal of the F-information matrix. This is another explanation for the optimistic standard error of the item parameters in JML estimation, which was also mentioned by Holland (1990, p.594).

The F-information matrices in the conditional distribution  $f$  and the marginal distribution  $g$  are given by

$$I_f(\beta; \omega) = \begin{pmatrix} .409 & -.272 & -.137 \\ -.272 & .451 & -.180 \\ -.137 & -.180 & .316 \end{pmatrix}$$

and

$$I_g(\beta; \omega) = \begin{pmatrix} .026 & .012 & -.037 \\ .012 & .021 & -.032 \\ -.037 & -.032 & .070 \end{pmatrix}$$

It is easily checked that all the F-information matrices are singular:

$$\begin{aligned} \Sigma_i(I_p(\beta; \omega))_{ij} &= \Sigma_j(I_p(\beta; \omega))_{ij} = \Sigma_i(I_f(\beta; \omega))_{ij} = \\ \Sigma_j(I_f(\beta; \omega))_{ij} &= \Sigma_i(I_g(\beta; \omega))_{ij} = \Sigma_j(I_g(\beta; \omega))_{ij} = 0 \end{aligned}$$

This is, as with the Fisher information matrix, due to the well known property that without any restriction the parameters are not identified in the Rasch model (Fischer & Molenaar, 1995). In order to compute the loss of information and the information efficiency for normalization the first item will be fixed. Information

matrices in which this normalization is applied, the first row and column are dropped, will be denoted by adding a star,  $\mathbf{I}^\star$ .

Computation of the information efficiency for using the conditional distribution  $f(\mathbf{x} | \mathbf{t}; \boldsymbol{\beta})$  instead of the full model  $p(\mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\theta})$  for estimating the item parameters proceeds in the example as follows:

$$\mathbf{I}_p^\star(\boldsymbol{\beta}; \boldsymbol{\omega}) = \begin{pmatrix} .472 & -.212 \\ -.212 & .386 \end{pmatrix} \text{ and } \det \mathbf{I}_p^\star(\boldsymbol{\beta}; \boldsymbol{\omega}) = 0.137$$

for the full model and for the conditional model we have:

$$\mathbf{I}_f^\star(\boldsymbol{\beta}; \boldsymbol{\omega}) = \begin{pmatrix} .451 & -.180 \\ -.180 & .316 \end{pmatrix} \text{ and } \det \mathbf{I}_f^\star(\boldsymbol{\beta}; \boldsymbol{\omega}) = 0.110.$$

The information efficiency using CML instead of JML estimation for the item parameters is then given by

$$\text{INFEFF}(\boldsymbol{\beta}; f:p) = \left( \frac{\det(\mathbf{I}_f^\star(\boldsymbol{\beta}; \boldsymbol{\omega}))}{\det(\mathbf{I}_p^\star(\boldsymbol{\beta}; \boldsymbol{\omega}))} \right)^{1/2} = (0.110/0.137)^{1/2} = .896.$$

In this example, the computation uses the normalization is on the first item. In chapter 3 of this dissertation it is shown that the value of the information efficiency  $\text{INFEFF}(\boldsymbol{\beta}; f:p)$  is independent of the chosen normalization.  $\square$

### 2.7.1 Comparison of information efficiency in JML and CML estimation

The efficiency of CML versus JML will be reported for some typical item parameter and ability parameters sets. For these sets, 100 ability parameters were drawn from a normal ability distribution  $N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ . Table 2.1 gives the information



efficiencies for 10 items with  $\beta = (-3, -2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2, 3)$  and 100 abilities drawn from the normal distribution with varying  $\mu$  and  $\sigma^2$ .

Table 2.1.  $(\det(\Gamma_f^*(\beta; \omega)))^{1/9}$ ,  $(\det(\Gamma_p^*(\beta; \omega)))^{1/9}$  and  $\text{INFEFF}(\beta; f; p)$  for  $\beta = (-3, -2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2, 3)$  and 100 abilities from  $N(\mu, \sigma^2)$

$\mu$	$\sigma^2$			
	.25	1	2.25	4
$(\det(\Gamma_f^*(\beta; \omega)))^{1/9}$				
-2	$5.38 \cdot 10^{-2}$	$5.80 \cdot 10^{-2}$	$6.27 \cdot 10^{-2}$	$6.56 \cdot 10^{-2}$
-1	$8.27 \cdot 10^{-2}$	$8.42 \cdot 10^{-2}$	$8.37 \cdot 10^{-2}$	$8.00 \cdot 10^{-2}$
0	$9.59 \cdot 10^{-2}$	$9.55 \cdot 10^{-2}$	$9.12 \cdot 10^{-2}$	$8.37 \cdot 10^{-2}$
1	$8.24 \cdot 10^{-2}$	$8.36 \cdot 10^{-2}$	$8.17 \cdot 10^{-2}$	$7.64 \cdot 10^{-2}$
2	$5.34 \cdot 10^{-2}$	$5.74 \cdot 10^{-2}$	$6.10 \cdot 10^{-2}$	$6.16 \cdot 10^{-2}$
$(\det(\Gamma_p^*(\beta; \omega)))^{1/9}$				
-2	$5.44 \cdot 10^{-2}$	$5.89 \cdot 10^{-2}$	$6.37 \cdot 10^{-2}$	$6.67 \cdot 10^{-2}$
-1	$8.38 \cdot 10^{-2}$	$8.56 \cdot 10^{-2}$	$8.52 \cdot 10^{-2}$	$8.16 \cdot 10^{-2}$
0	$9.73 \cdot 10^{-2}$	$9.71 \cdot 10^{-2}$	$9.30 \cdot 10^{-2}$	$8.54 \cdot 10^{-2}$
1	$8.35 \cdot 10^{-2}$	$8.49 \cdot 10^{-2}$	$8.32 \cdot 10^{-2}$	$7.79 \cdot 10^{-2}$
2	$5.40 \cdot 10^{-2}$	$5.82 \cdot 10^{-2}$	$6.20 \cdot 10^{-2}$	$6.27 \cdot 10^{-2}$
$\text{INFEFF}(\beta; f; p)$				
-2	.988	.986	.984	.983
-1	.986	.984	.982	.981
0	.986	.984	.981	.980
1	.987	.984	.982	.981
2	.988	.986	.983	.982

It can be noted that the determinants of  $\Gamma_f^*(\beta; \omega)$  as well as of  $\Gamma_p^*(\beta; \omega)$  decrease when the distance between the mean of the ability- and the mean of the item parameters,  $|\mu - \frac{1}{k} \sum_k \beta_i|$ , is increased. Furthermore, we see that for extreme  $\mu$  the determinants are increasing with  $\sigma^2$ . With respect to the information efficiency of CML versus JML, the results in Table 2.1, hardly show any variation. It is seen that the efficiency is at least 98.0%.

Table 2.2 shows the relation between the information efficiency and the spread in the item parameters. For 100 abilities drawn from  $N(0,1)$ , the efficiency for estimating 10 equidistant item parameters with  $\sum_i \beta_i = 0$  and varying stepsize,  $\beta_{i+1} - \beta_i$ ,  $i = 1, \dots, 9$  is given.

Table 2.2.  $\text{INFEFF}(\beta; f; p)$  for 10 equidistant items with  $\sum_i \beta_i = 0$  and varying stepsize and 100 abilities from  $N(0,1)$ .

step	$\text{INFEFF}(\beta; f; p)$
1	.951
.5	.986
.25	.996
.125	.999
.0625	1.000
0	1.000

The information efficiency of CML versus JML estimation increases with decreasing spread of the item parameters. In case all item parameters are equal there is no loss of information. The theoretical explanation for this is that the weak ancillarity condition in Theorem 3 (23) is fulfilled in this special case:  $\partial \ln \prod_v \gamma_{t_v}(\beta) / \partial \beta_j = \sum_v t_v / k$  for  $j = 1, \dots, k$  and  $I_g(\beta; \omega) = 0$ . (See chapter 3 of this thesis).

In Table 2.3 the relation between the information efficiency and the test length is illustrated. For 100 abilities drawn from  $N(0.5, 1)$ , the efficiency is reported for estimating the item parameters  $\beta = (0, 1, 2)$ , which are  $n$  times in a test. The test length is equal to  $3n$ .

Table 2.3.  $\text{INFEFF}(\beta; f; p)$  for  $\beta = (0, 1, 2)$  with varying test length and 100 abilities from  $N(0.5, 1)$

length	$\text{INFEFF}(\beta; f; p)$
3	.925
6	.989
9	.996
12	.998
15	.999
18	.999
21	.999

An increasing efficiency is observed when the test length is increased. It is clear that already at a relative short test length of 15 items, CML estimation of the item parameters provides almost the same amount of information as JML estimation.

It should be noted that the above results only show that there is loss in information if CML is used instead of JML and that this loss is small. This does not imply of course that JML is to be preferred above CML estimation for estimating the item parameters. On the contrary, the consistency and sampling independence of the CML estimators are stronger properties for the preference of CML above JML than the small loss of information.

## 2.8 F-information in the dichotomous Rasch model: comparing MML and CML

In order to compare CML with MML estimation of the item parameters in the Rasch model, the marginal likelihood function (3) in which the  $\theta_v$ 's are not fixed ability parameters but random draws from the ability distribution  $h(\theta; \xi)$ , is considered. With again  $T_v = \sum_i X_{vi}$ , this can be rewritten as:

$$p_m(x; \beta, \xi) = \prod_v p_m(x_v; \beta, \xi) = \prod_v \int \prod_i P(X_{vi} = x_{vi} | \theta_v; \beta_i) h(\theta_v; \xi) d\theta_v = \prod_v f(x_v | t_v; \beta) \prod_v \int g(t_v | \theta_v; \beta) h(\theta_v; \xi) d\theta_v. \quad (32)$$

In (32) the parameters are  $\beta$  the item parameter and  $\xi$  the parameter of the ability distribution.  $p_m(x; \beta, \xi)$  is factored in a part which is only dependent on the parameter of interest  $\beta$ , and a part which is the distribution of  $T$ , which depends on both parameters  $\beta$  and  $\xi$ . For estimating the item parameters with CML, as seen in (5), only the first part of the factorization is used. In MML estimation the full model (32) is used. And for comparing CML estimation with MML estimation, the F-information in  $f$ ,  $\mathcal{E} I_f(\beta | T)$ , given in (31ab), can be compared with the F-information in  $p_m$  with respect to  $\beta$ .

First note that because in the model (32) every person is a random draw from an ability distribution, it suffices to derive the F-information in  $p_m$  for only one person, indexed with  $v$ . The same is true for the F-information in  $f$  (see (31ab)).

In order to determine these F-information matrices completely, an ability distribution  $h(\theta_v; \xi)$  has to be specified. Here it is assumed that the ability distribution is normal with parameter  $\xi = (\mu, \sigma^2)$ , and the density is given by

$$h(\theta_v; \xi) = \sigma^{-1} \phi((\theta_v - \mu)/\sigma), \quad (33)$$

with  $\phi(y)$  the standard normal density:

$$\phi(y) = (2\pi)^{-1/2} \exp(-y^2/2). \quad (34)$$

### 2.8.1 F-information in $p_m$

As before, this matrix can be determined once the Fisher information matrix has been obtained. If the length of the vector parameter of the ability distribution is  $\ell$ , this matrix can be written in the following partitioned form:

$$I_{p_m}(\omega) = \mathcal{E} \begin{bmatrix} S_{p;\beta} & S_{p;\beta}^T & S_{p;\beta} & S_{p;\xi}^T \\ S_{p;\xi} & S_{p;\beta}^T & S_{p;\xi} & S_{p;\xi}^T \end{bmatrix} = : \begin{bmatrix} I_{p_m}^{\beta\beta^T} & I_{p_m}^{\beta\xi^T} \\ I_{p_m}^{\xi\beta^T} & I_{p_m}^{\xi\xi^T} \end{bmatrix} \quad (35)$$

The submatrices,  $I_{p_m}^{\beta\beta^T} (k \times k)$ ,  $I_{p_m}^{\beta\xi^T} = [I_{p_m}^{\xi\beta^T}]^T (k \times \ell)$  and  $I_{p_m}^{\xi\xi^T} (\ell \times \ell)$  are again the negative expectation of the second derivatives of  $\ln p_m$  with respect to the parameters. For notational convenience define:

$$Q_v := Q_v(\theta_v, x_{vi}, t_v, \beta, \mu, \sigma^2) = \exp(\theta_v t_v) \prod_{i=1}^k (1 - p_{vi}) h(\theta_v; \xi), \quad (36)$$

and the normal density ((33) with (34)), can be substituted for  $h(\theta_v; \xi)$ .

Consider the model for one person  $v$ :

$$p_m(x_v; \beta, \xi) = f(x_v | t_v; \beta) \cdot \int_{-\infty}^{\infty} g(t_v | \theta_v; \beta) h(\theta_v; \xi) d\theta_v. \quad (37)$$

In (37)  $g(t_v | \theta_v; \beta)$ , referring to (22) and using the notational definition in (25) and (36), is

$$g(t_v | \theta_v; \beta) = \exp(\theta_v t_v) \cdot \gamma_{t_v}(\beta) \cdot \prod_{i=1}^k (1 - p_{vi}) = Q_v \gamma_{t_v}(\beta) / h(\theta_v; \xi), \quad (38)$$

and

$$f(x_v | t_v; \beta) = \frac{\exp(-\sum_i \beta_i x_{vi})}{\gamma_{t_v}(\beta)}, \quad (39)$$

which can be found by dividing the expressions (21) and (22) for one person.

The marginal distribution of  $T_v$  is given by

$$P(T_v = t_v; \beta, \xi) = \int_{-\infty}^{\infty} g(t_v | \theta_v; \beta) h(\theta_v; \xi) d\theta_v = \int Q_v d\theta_v. \quad (40)$$

Inserting (38) and (39) in (37) and taking the logarithm, gives

$$\ln p_m(x_v; \beta, \xi) = -\sum_i \beta_i x_{vi} + \ln \int_{-\infty}^{\infty} \exp(\theta_v t_v) h(\theta_v; \xi) \prod_{i=1}^k (1 - p_{vi}) d\theta_v = -\sum_i \beta_i x_{vi} + \ln \int Q_v d\theta_v$$

With some algebra the following expressions are easily derived:

$$\begin{aligned} \frac{\partial Q_v}{\partial \beta_j} &= Q_v p_{vj}, \text{ for } j = 1, \dots, k, \\ \frac{\partial Q_v}{\partial \mu} &= Q_v (\theta_v - \mu) \sigma^{-2}, \\ \frac{\partial Q_v}{\partial \sigma^2} &= Q_v \left[ -\frac{1}{2} \sigma^{-2} + \frac{1}{2} (\theta_v - \mu)^2 \sigma^{-4} \right], \\ \frac{\partial^2 Q_v}{\partial \beta_j^2} &= Q_v (2p_{vj} - 1) p_{vj}, \text{ for } j = 1, \dots, k, \end{aligned} \quad (41a)$$

$$\frac{\partial^2 Q_v}{\partial \beta_j \partial \beta_m} = Q_v p_{vj} p_{vm}, \text{ for } m \neq j = 1, \dots, k, \quad (41b)$$

$$\frac{\partial^2 Q_v}{\partial \mu^2} = Q_v [(\theta_v - \mu)^2 \sigma^{-4} - \sigma^{-2}], \quad (42a)$$

$$\frac{\partial^2 Q_v}{\partial (\sigma^2)^2} = Q_v \left[ \frac{3}{4} \sigma^{-4} - \frac{3}{2} (\theta_v - \mu)^2 \sigma^{-6} + \frac{1}{4} (\theta_v - \mu)^4 \sigma^{-8} \right] \quad (42b)$$

$$\frac{\partial^2 Q_v}{\partial \sigma^2 \partial \mu} = Q_v \left[ -\frac{3}{2} (\theta_v - \mu) \sigma^{-4} + \frac{1}{2} (\theta_v - \mu)^3 \sigma^{-6} \right], \quad (42c)$$

$$\frac{\partial^2 Q_v}{\partial \mu \partial \beta_j} = Q_v p_{vj} (\theta_v - \mu) \sigma^{-2}, \text{ for } j = 1, \dots, k, \quad (43a)$$

$$\frac{\partial^2 Q_v}{\partial \sigma^2 \partial \beta_j} = Q_v p_{vj} \left[ -\frac{1}{2} \sigma^{-2} + \frac{1}{2} (\theta_v - \mu)^2 \sigma^{-4} \right] \text{ for } j = 1, \dots, k. \quad (43b)$$

The second derivatives of  $\ln p_m(x_v; \beta, \xi)$ <sup>1</sup> are generally written as:

$$\frac{\partial^2 \ln p_m(x_v; \beta, \xi)}{\partial a \partial b} = \frac{\int \frac{\partial^2 Q_v}{\partial a \partial b} d\theta_v}{\int Q_v d\theta_v} - \frac{\int \frac{\partial Q_v}{\partial a} d\theta_v \int \frac{\partial Q_v}{\partial b} d\theta_v}{(\int Q_v d\theta_v)^2}, \quad (44)$$

with  $a$  and  $b$  being any pair from the parameters  $\mu$ ,  $\sigma^2$ , and  $\beta_j, j = 1, \dots, k$ . Taking minus the expectation of (44) over the distribution of  $T_v$  (40) gives:

$$\mathcal{E} \frac{\partial^2 \ln p_m(x_v; \beta, \xi)}{\partial a \partial b} = - \sum_{t_v=0}^k \gamma_{t_v} \left( \int \frac{\partial^2 Q_v}{\partial a \partial b} d\theta_v - \frac{\int \frac{\partial Q_v}{\partial a} d\theta_v \int \frac{\partial Q_v}{\partial b} d\theta_v}{\int Q_v d\theta_v} \right), \quad (45)$$

which is the general expression for all parts of the Fisher information matrix (35).

Using (41ab) in (45) gives the item parameter part:

$$(\mathbf{I}_{P_m}^{\beta\beta^T})_{jj} = - \sum_{t_v=0}^k \gamma_{t_v} \left[ \int (2p_{vj} - 1) p_{vj} Q_v d\theta_v - \frac{(\int p_{vj} Q_v d\theta_v)^2}{\int Q_v d\theta_v} \right] \quad (46a)$$

for  $j = 1, \dots, k$ , and

$$(\mathbf{I}_{P_m}^{\beta\beta^T})_{mj} = - \sum_{t_v=0}^k \gamma_{t_v} \left[ \int p_{vm} p_{vj} Q_v d\theta_v - \frac{\int p_{vm} Q_v d\theta_v \int p_{vj} Q_v d\theta_v}{\int Q_v d\theta_v} \right] \quad (46b)$$

for  $m \neq j = 1, \dots, k$ .

Using (42abc) in (45) gives the ability distribution parameter part:

$$(\mathbf{I}_{P_m}^{\xi\xi^T})_{11} = - \sum_{t_v=0}^k \gamma_{t_v} \left( \int Q_v [(\theta_v - \mu)^2 \sigma^{-4} - \sigma^{-2}] d\theta_v - \frac{(\int Q_v (\theta_v - \mu) \sigma^{-2} d\theta_v)^2}{\int Q_v d\theta_v} \right) \quad (47a)$$

$$(\mathbf{I}_{P_m}^{\xi\xi^T})_{22} = - \sum_{t_v=0}^k \gamma_{t_v} \left\{ \int Q_v \left[ \frac{3}{4} \sigma^{-4} - \frac{3}{2} (\theta_v - \mu)^2 \sigma^{-6} + \frac{1}{4} (\theta_v - \mu)^4 \sigma^{-8} \right] d\theta_v - \frac{\left( \int Q_v \left[ -\frac{1}{2} \sigma^{-2} + \frac{1}{2} (\theta_v - \mu)^2 \sigma^{-4} \right] d\theta_v \right)^2}{\int Q_v d\theta_v} \right\} \quad (47b)$$

$$(\mathbf{I}_{P_m}^{\xi\xi^T})_{12} = - \sum_{t_v=0}^k \gamma_{t_v} \left\{ \int Q_v \left[ -\frac{3}{2} (\theta_v - \mu) \sigma^{-4} + \frac{1}{2} (\theta_v - \mu)^3 \sigma^{-6} \right] d\theta_v - \frac{\int Q_v (\theta_v - \mu) \sigma^{-2} d\theta_v \int Q_v \left[ -\frac{1}{2} \sigma^{-2} + \frac{1}{2} (\theta_v - \mu)^2 \sigma^{-4} \right] d\theta_v}{\int Q_v d\theta_v} \right\}. \quad (47c)$$

Finally, using (43ab) in (45) gives the elements of  $\mathbf{I}_{P_m}^{\beta\xi^T}$ . For  $j=1,...,k$

$$(\mathbf{I}_{P_m}^{\beta\xi^T})_{1j} = - \sum_{t_v=0}^k \gamma_{t_v} \left\{ \int Q_v p_{vj} (\theta_v - \mu) \sigma^{-2} d\theta_v - \frac{\int Q_v p_{vj} d\theta_v \int Q_v (\theta_v - \mu) \sigma^{-2} d\theta_v}{\int Q_v d\theta_v} \right\} \quad (48a)$$

and

$$(\mathbf{I}_{P_m}^{\beta\xi^T})_{2j} = - \sum_{t_v=0}^k \gamma_{t_v} \left\{ \int Q_v p_{vj} \left[ -\frac{1}{2} \sigma^{-2} + \frac{1}{2} (\theta_v - \mu)^2 \sigma^{-4} \right] d\theta_v - \frac{\int Q_v p_{vj} d\theta_v \int Q_v \left[ -\frac{1}{2} \sigma^{-2} + \frac{1}{2} (\theta_v - \mu)^2 \sigma^{-4} \right] d\theta_v}{\int Q_v d\theta_v} \right\}. \quad (48b)$$

With respectively (46ab), (47abc), and (48ab), all expressions for the submatrices of the Fisher information matrix (35) are obtained. From these expressions the F-information matrix  $\mathbf{I}_{p_m}(\boldsymbol{\beta}; \omega)$  can be computed. Because  $\mathbf{I}_{p_m}^{\xi\xi^T}$  is positive definite, the F-information matrix with respect to  $\boldsymbol{\beta}$  is given by (13)

$$\mathbf{I}_{p_m}(\boldsymbol{\beta}; \omega) = \mathbf{I}_{p_m}^{\beta\beta^T} - \mathbf{I}_{p_m}^{\beta\xi^T} [\mathbf{I}_{p_m}^{\xi\xi^T}]^{-1} \mathbf{I}_{p_m}^{\xi\beta^T}.$$

### 2.8.2 F-information in $f(x|t)$

In order to make the relevant comparisons, the expressions for the F-information in the conditional distribution are required. Taking expectations over the distribution of  $T_v$  (40) in (31ab) gives:

$$(\mathbf{I}_f(\boldsymbol{\beta}; \omega))_{jj} = (\mathcal{E} \mathbf{I}_f(\boldsymbol{\beta} | T))_{jj} = \sum_{t_v=0}^k \left\{ e^{-\beta_j \cdot \gamma_{t_v-1}^{(j)}} - \frac{\left[ e^{-\beta_j \cdot \gamma_{t_v-1}^{(j)}} \right]^2}{\gamma_{t_v}} \right\} \int Q_v d\theta_v$$

for  $j = 1, \dots, k$ , and

$$(\mathbf{I}_f(\boldsymbol{\beta}; \omega))_{mj} = (\mathcal{E} \mathbf{I}_f(\boldsymbol{\beta} | T))_{mj} = \sum_{t_v=0}^k \left\{ - \frac{e^{-\beta_m - \beta_j} \cdot \gamma_{t_v-1}^{(m)} \gamma_{t_v-1}^{(j)}}{\gamma_{t_v}} + e^{-\beta_m - \beta_j} \cdot \gamma_{t_v-2}^{(mj)} \right\} \int Q_v d\theta_v$$

for  $m \neq j = 1, \dots, k$ .

#### Example

Consider the example similar to the one used in the computations of the information matrices in the Rasch model with fixed ability parameters. For the Rasch model with 3 items,  $\beta_1 = -.5$ ,  $\beta_2 = 0.0$  and  $\beta_3 = 1.75$ , and assuming that  $\boldsymbol{\theta}$  is distributed as  $N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ , with  $\boldsymbol{\mu} = \mathbf{0}$  and  $\boldsymbol{\sigma}^2 = 1$ , the expected Fisher information matrix,  $(k + \ell) \times (k + \ell)$  (35), for one random draw of the ability distribution is given by:



$$I_{p_m}(\omega) = \begin{pmatrix} I_{p_m}^{\beta\beta^T} & I_{p_m}^{\beta\xi^T} \\ I_{p_m}^{\xi\beta^T} & I_{p_m}^{\xi\xi^T} \end{pmatrix} = \begin{pmatrix} .172 & -.027 & -.016 & -.129 & .016 \\ -.027 & .178 & -.018 & -.134 & .002 \\ -.016 & -.018 & .119 & -.085 & -.029 \\ -.129 & -.134 & -.085 & .348 & .011 \\ .016 & .002 & -.029 & .011 & .038 \end{pmatrix}$$

As expected, the upper left (3×3) submatrix  $I_{p_m}^{\beta\beta^T}$  is symmetric, and the off diagonal elements, the covariances between the score statistics with respect to different items, are negative. This submatrix is the information matrix with respect to  $\beta$ , when  $\mu$  and  $\sigma^2$  are considered known. In the submatrix  $I_{p_m}^{\beta\xi^T}$ , negative covariances between the score statistics of item parameters and of the mean of the ability distribution are obtained, while these covariances for item parameters and the variance of the ability distribution has no consistent sign. In the part of the parameters of the ability distribution, a non-zero covariance between the score statistics of  $\mu$  and  $\sigma^2$  can be observed. Note that only in case the mean of the item parameters equals the mean of the ability distribution, this covariance is zero.

The F-Information matrices in  $p_m$  and in the conditional distribution  $f$  are given by

$$I_{p_m}(\beta; \omega) = \begin{pmatrix} .114 & -.080 & -.034 \\ -.080 & .126 & -.046 \\ -.034 & -.046 & .080 \end{pmatrix} \text{ and}$$

$$I_f(\beta; \omega) = \mathcal{E} I_f(\beta | T) = \begin{pmatrix} .114 & -.080 & -.034 \\ -.080 & .122 & -.042 \\ -.034 & -.042 & .076 \end{pmatrix}$$

As before these matrices are singular and for normalization the first item will be fixed. The information efficiency for estimating the item parameters  $\beta$  using the

conditional model (CML) instead of the full model (MML) is then computed as follows:

$$I_{p_m}^*(\beta; \omega) = \begin{pmatrix} .126 & -.046 \\ -.046 & .080 \end{pmatrix} \text{ and } \det I_{p_m}^*(\beta; \omega) = 0.00794$$

for MML and for CML we have:

$$I_f^*(\beta; \omega) = \begin{pmatrix} .122 & -.042 \\ -.042 & .076 \end{pmatrix} \text{ and } \det I_f^*(\beta; \omega) = 0.00747.$$

The information efficiency using CML instead of JML estimation for the item parameters is then given by

$$\text{INFEFF}(\beta; f; p) = \left( \frac{\det(I_f^*(\beta; \omega))}{\det(I_{p_m}^*(\beta; \omega))} \right)^{1/2} = (0.00747/0.00794)^{1/2} = .970 \square$$

### 2.8.3 Comparison of information efficiency in CML and MML estimation

The information efficiency of CML versus MML will be reported for some typical item parameter sets. In Table 2.4, the determinants of the F-information matrices and the efficiency for 10 items with  $\beta = (-3, -2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2, 3)$  with varying  $\mu$  and  $\sigma^2$  are presented.

It can be seen that the determinants of the F-information matrices  $I_f^*(\beta; \omega)$  as well as  $I_{p_m}^*(\beta; \omega)$ , and, as a consequence of that, also the efficiencies, are symmetric in  $\mu$ . Furthermore it is seen that for extreme  $\mu$  the determinants are increasing with increasing  $\sigma^2$ , while for  $\mu = 0$  the determinants decrease with increasing  $\sigma^2$ . Finally it is observed, that the determinants decrease when the distance between the mean of the ability- and the mean of the item parameters,  $|\mu - \sum_i \beta_i / k|$ , is increased. It seems that the more the abilities are closer to the item parameters, the determinants of the information matrices are larger.

Table 2.4.  $(\det(I_f^*(\beta; \omega)))^{1/9}$ ,  $(\det(I_{p_m}^*(\beta; \omega)))^{1/9}$  and  $\text{INFEFF}(\beta; f; p_m)$  for  $\beta = (-3, -2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2, 3)$  and ability from  $N(\mu, \sigma^2)$

$\mu$	$\sigma^2$			
	.25	1	2.25	4
	$(\det(I_f^*(\beta; \omega)))^{1/9}$			
-2	5.36 $10^{-2}$	5.78 $10^{-2}$	6.17 $10^{-2}$	6.30 $10^{-2}$
-1	8.25 $10^{-2}$	8.38 $10^{-2}$	8.24 $10^{-2}$	7.83 $10^{-2}$
0	9.59 $10^{-2}$	9.52 $10^{-2}$	9.10 $10^{-2}$	8.43 $10^{-2}$
1	8.25 $10^{-2}$	8.38 $10^{-2}$	8.24 $10^{-2}$	7.83 $10^{-2}$
2	5.36 $10^{-2}$	5.78 $10^{-2}$	6.17 $10^{-2}$	6.30 $10^{-2}$
	$(\det(I_{p_m}^*(\beta; \omega)))^{1/9}$			
-2	5.37 $10^{-2}$	5.82 $10^{-2}$	6.23 $10^{-2}$	6.39 $10^{-2}$
-1	8.28 $10^{-2}$	8.84 $10^{-2}$	8.35 $10^{-2}$	7.96 $10^{-2}$
0	9.61 $10^{-2}$	9.60 $10^{-2}$	9.22 $10^{-2}$	8.56 $10^{-2}$
1	8.28 $10^{-2}$	8.84 $10^{-2}$	8.35 $10^{-2}$	7.96 $10^{-2}$
2	5.37 $10^{-2}$	5.82 $10^{-2}$	6.23 $10^{-2}$	6.39 $10^{-2}$
	$\text{INFEFF}(\beta; f; p_m)$			
-2	.9969	.9935	.9892	.9857
-1	.9967	.9922	.9875	.9845
0	.9969	.9919	.9870	.9841
1	.9967	.9922	.9875	.9845
2	.9969	.9935	.9892	.9857

The information efficiency of CML versus MML decreases somewhat with increasing  $\sigma^2$ . However, the efficiency of CML versus MML amounts to at least .9841.

In Table 2.5, the relation between the information efficiency and the spread in the item parameters is given. For 1 ability drawn from  $N(0,1)$ , the efficiency for estimating 10 equidistant item parameters with  $\sum_i \beta_i = 0$  and varying stepsize,  $\beta_{i+1} - \beta_i$ ,  $i = 1, \dots, 9$  is given.

Table 2.5.  $\text{INFEFF}(\beta; f; p_m)$  for  $\sum_i \beta_i = 0$  with varying stepsize and 1 ability from  $N(0,1)$ 

step	$\text{INFEFF}(\beta; f; p_m)$
1	.9719
.5	.9943
.25	.9993
.125	.9999
.0625	1.0000

Although there is hardly any loss in information, the efficiency of CML versus MML estimation increases with a decrease in the spread between the item parameters.

In Table 2.6, the relation between the information efficiency and the test length is illustrated. For 1 ability drawn from  $N(0.5,1)$ , the efficiency is reported for estimating the item parameters  $\beta = (0,1,2)$ , which are  $n$  times in a test. The test length is  $3n$ .

Table 2.6.  $\text{INFEFF}(\beta; f; p_m)$  for  $\beta_i = 0,1,2$  with varying test length and 1 ability from  $N(0.5,1)$ 

length	$\text{INFEFF}(\beta; f; p_m)$
3	.9842
6	.9984
9	.9993
12	.9996
15	.9997
18	.9998

When the test length is increased, the information efficiency also increases. Already with very short tests CML estimation is almost as efficient in the use of information as MML estimation of the item parameters.

## **2.9 Conclusion**

In this study it is shown that the concept of F-information, a generalization of Fisher information, is a useful tool for evaluating the loss of information in conditional maximum likelihood estimation. In separable models with a sufficient statistic for the nuisance parameter, the conditional model which only depends on the parameter of interest is often used instead of the full model for estimating the parameter of interest. With the F-information concept it is possible to investigate the conditions under which there is no loss of information in CML estimation, and furthermore, if there is a loss, to quantify this.

In the chapter the main properties of F-information are presented and the conditions for no loss of information are specified. Especially in the case of exponential family models, it is shown that these conditions can be easily checked. It is shown that in the Poisson Counts Model these conditions are met. For the Rasch model for dichotomously scored items these conditions are not fulfilled, which means that loss of information in CML estimation of the item parameters in this model is to be expected.

For the dichotomous Rasch model the expressions needed to be able to investigate the loss of information using CML estimation of the item parameters are derived in detail. For the Rasch model with fixed ability parameters, a comparison was made between JML and CML estimation, and, under the assumption of a normal ability distribution, a comparison of MML and CML estimation. The comparisons are made in several conditions. Varied were: the spread of the item difficulty parameters, the mean and the variance of the ability distribution and the test length. In almost all comparisons, some loss of information in using CML appeared. However, in all the comparisons of CML to JML as well as to MML the loss showed to be very small. The reported information efficiencies are always larger than 92%, and in the comparison of CML versus MML estimation even larger than 97%.

On basis of these results it can be concluded that CML item parameter estimation in the Rasch model, which has some other known practical and theoretical attractive properties, is also from an information point of view a sound

practice. Hardly any loss of information is to be expected compared to alternative estimation methods.

The method of information comparison described in the paper, applies to any separable model. For popular extensions of the dichotomous Rasch model (Fischer & Molenaar, 1995), for example the partial credit model and the one-parameter logistic model, the methods described in this chapter can be generalized and applied.

## 2.10 References

- Andersen, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32, 283-301
- Andersen, E.B. (1973). *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk Forlag.
- Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Society*, 72, 355-366.
- Bhapkar, V.P. (1989). Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *Journal of Statistical Planning and Inference*, 21, 139-160.
- Bhapkar, V.P. (1991). Loss of information in the presence of nuisance parameters and partial sufficiency. *Journal of Statistical Planning and Inference*, 28, 185-203.
- De Leeuw, J. & Verhelst, N.D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics*, 11, 183-196.
- Efron, B. (1977). On the efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Society*, 72, 557-565.
- Engelen, R.J.H. (1989). *Parameter estimation in the logistic item response model*. (Doctoral Thesis.) Enschede: University of Twente.
- Fischer, G.H. (1974). *Einführung in die Theorie der psychologischer Tests*. [Introduction to mental test theory.] Bern: Huber.
- Fischer, G.H., & Molenaar, I.W. (Eds.) (1995). *Rasch Models*. New York: Springer.
- Glas, C.A.W. (1989). Contributions to estimating and testing Rasch models. (Doctoral Thesis.) Enschede: University of Twente.
- Holland, P.W. (1990). On the sampling theory foundations of item response models. *Psychometrika*, 55, 577-601.
- Liang, K. (1983). On information and ancillary in the presence of a nuisance parameter. *Biometrika*, 70, 607-612.

- Lindsay, B., Clogg, C.C. & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86, 96-107.
- Louis, T.A. (1982). Finding the observed information matrix with the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44, 226-233.
- Molenaar, I.W. (1995). Estimation of item parameters. In: Fischer, G.H. & Molenaar, I.W. (Eds.). *Rasch Models* (pp.39-51). New York: Springer.
- Pfanzagl, J. (1993). A case of asymptotic equivalence between conditional and marginal maximum likelihood estimators. *Journal of Statistical Planning and Inference*, 35, 301-307.
- Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In G.H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 249-263). New York: Springer Verlag.
- Pukelsheim, F. (1993). *Optimal design of experiments*. New York: Wiley.
- Rao, C.R. (1973). *Linear statistical inference and its applications*. New York: Wiley.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)
- Wright, B.J. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97-116.



## 1. Note

If the ability distribution  $h(\theta_v; \xi)$  is considered as prior distribution, the a posteriori distribution of the ability of person  $v$ ,  $\theta_v$  given the score  $t_v$ , is by Bayes rule equal to:

$$k(\theta_v | t_v; \beta, \xi) = : \frac{g(t_v | \theta_v; \beta) h(\theta_v; \xi)}{\int_{-\infty}^{\infty} g(t_v | \theta_v; \beta) h(\theta_v; \xi) d\theta_v} = \frac{Q_v}{\int Q_v d\theta_v}.$$

The second derivatives with respect to the item parameters can be rewritten as:

$$\frac{\partial^2 \ln p_m(x_v; \beta, \xi)}{\partial \beta_j^2} = \int_{-\infty}^{\infty} (2p_{vj} - 1) p_{vj} k(\theta_v | t_v; \beta, \xi) d\theta_v - \left[ \int_{-\infty}^{\infty} p_{vj} k(\theta_v | t_v; \beta, \xi) d\theta_v \right]^2 \quad (i)$$

for  $j = 1, \dots, k$ ; and

$$\begin{aligned} \frac{\partial^2 \ln p_m(x_v; \beta, \xi)}{\partial \beta_m \partial \beta_j} &= \int_{-\infty}^{\infty} p_{vm} p_{vj} k(\theta_v | t_v; \beta, \xi) d\theta_v - \\ &\left[ \int_{-\infty}^{\infty} p_{vm} k(\theta_v | t_v; \beta, \xi) d\theta_v \right] \left[ \int_{-\infty}^{\infty} p_{vj} k(\theta_v | t_v; \beta, \xi) d\theta_v \right] \end{aligned} \quad (ii)$$

for  $m \neq j = 1, \dots, k$ .

(i) and (ii) can be written as moments in the a posteriori distribution of  $\theta_v$  given  $t_v$ . Because  $(2p_{vj} - 1)p_{vj} = -(1 - p_{vj})p_{vj} + p_{vj}^2$ , it yields that:

$$\begin{aligned} \frac{\partial^2 \ln p_m(x_v; \beta, \xi)}{\partial \beta_j^2} &= -\mathcal{E}(p_{vj}(1 - p_{vj}) | t_v) + \text{Var}(p_{vj} | t_v) \\ \frac{\partial^2 \ln p_m(x_v; \beta, \xi)}{\partial \beta_m \partial \beta_j} &= \text{Cov}(p_{vm}, p_{vj} | t_v). \end{aligned}$$

Evaluating the negatives of these expressions in the estimates of the parameters, the item parameter part of the observed information matrix of the Rasch model using MML estimation are obtained. The same expressions were deduced by Glas (1989, p.50-55, Appendix D), using the theory of Louis (1982) on the relation between the observed and expected information matrix.

## Appendix chapter 2

### Proof Property 4a

$$\begin{aligned}\mathcal{E}(S_{p;\Psi} | T=t) &= \int_{\{x: t(x)=t\}} \frac{\partial \ln p(x;\omega)}{\partial \Psi} \cdot f(x | t;\omega) dx = \int_{\{x: t(x)=t\}} \frac{\partial p(x;\omega)/\partial \Psi}{p(x;\omega)} \cdot \frac{p(x;\omega)}{g(t;\omega)} dx = \\ &= \int_{\{x: t(x)=t\}} \frac{\partial p(x;\omega)/\partial \Psi}{g(t;\omega)} dx = \frac{\partial g(t;\omega)/\partial \Psi}{g(t;\omega)} = \frac{\partial \ln g(t;\omega)}{\partial \Psi} = S_{g;\Psi}\end{aligned}$$

The fourth equality holds because  $g(t;\omega) = \int_{\{x: t(x)=t\}} p(x;\omega) dx$ .  $\square$

### Proof Property 5a

Following a general property of conditional expectations

$$\mathcal{E}(S_{p;\Psi} S_{g;\Psi}^T) = \mathcal{E}_T(\mathcal{E}_{X|T}(S_{p;\Psi} S_{g;\Psi}^T | T=t)),$$

but in the conditional distribution  $S_{g;\Psi}$  is a constant. So

$$\mathcal{E}_{X|T}(S_{p;\Psi} S_{g;\Psi}^T | T=t) = \mathcal{E}_{X|T}(S_{p;\Psi} | T=t) S_{g;\Psi}^T.$$

Then using property 4a, 5a follows.  $\square$

### Proof Property 7a

Because of property 3 we have:  $\mathcal{E}(S_{p;\Psi} S_{p;\tau}^T + S_{p;\Psi,\tau} S_{p;\tau}^T) = \mathcal{E}(S_{p;\Psi} S_{p;\tau}^T + \frac{\partial}{\partial \Psi} S_{p;\tau} S_{p;\tau}^T) = 0$ .

Applying property 6 this can be rewritten as:

$$\mathcal{E}[(S_{f;\Psi} + S_{g;\Psi}) S_{g;\tau}^T + \frac{\partial}{\partial \Psi} S_{p;\tau} S_{p;\tau}^T] = 0,$$

$$\mathcal{E}[S_{f;\Psi} S_{g;\tau}^T + S_{g;\Psi} S_{g;\tau}^T + \frac{\partial}{\partial \Psi} S_{g;\tau} S_{g;\tau}^T] = 0,$$

$$\mathcal{E}(S_{f;\Psi} S_{g;\tau}^T) + \mathcal{E}(S_{g;\Psi} S_{g;\tau}^T) + \mathcal{E} S_{g;\Psi,\tau} S_{g;\tau}^T = 0.$$

And the last two terms in the left hand side of the last equation sum to 0, because of property 3a.  $\square$

*Proof Property 8a*

Because of property 5a

$$\mathcal{E} (S_{p;\psi} S_{g;\psi}^T - S_{g;\psi} S_{g;\psi}^T) = 0,$$

$$\mathcal{E} [(S_{p;\psi} - S_{g;\psi}) S_{g;\psi}^T] = 0.$$

Using property 6 this becomes

$$\mathcal{E} [(S_{f;\psi} + S_{g;\psi} - S_{g;\psi}) S_{g;\psi}^T] = \mathcal{E} (S_{f;\psi} S_{g;\psi}^T) = 0. \square$$

*Proof Theorem 1:*

a. Using (12) and (13) we write

$$I_p(\psi; \omega) = \mathcal{E} [S_{p;\psi} S_{p;\psi}^T] - \mathcal{E} [S_{p;\psi} S_{p;\tau}^T] (\mathcal{E} [S_{p;\tau} S_{p;\tau}^T])^{-1} \mathcal{E} [S_{p;\tau} S_{p;\psi}^T]. \quad (49)$$

Replacing in (49) the efficient score statistics in  $p$  by those in  $f$  and  $g$ , using property 6 above, gives:

$$I_p(\psi; \omega) = \mathcal{E} (S_{f;\psi} S_{f;\psi}^T + S_{f;\psi} S_{g;\psi}^T + S_{g;\psi} S_{f;\psi}^T + S_{g;\psi} S_{g;\psi}^T) - \mathcal{E} (S_{f;\psi} S_{g;\tau}^T + S_{g;\psi} S_{g;\tau}^T) (\mathcal{E} [S_{g;\tau} S_{g;\tau}^T])^{-1} \mathcal{E} (S_{g;\tau} S_{f;\psi}^T + S_{g;\tau} S_{g;\psi}^T).$$

In this expression, due to properties 7 and 8,

$$\mathcal{E} (S_{f;\psi} S_{g;\psi}^T) = \mathcal{E} (S_{g;\psi} S_{f;\psi}^T) = \mathcal{E} (S_{f;\psi} S_{g;\tau}^T) = \mathcal{E} (S_{g;\tau} S_{f;\psi}^T) = 0. \quad \text{So,}$$

$$I_p(\psi; \omega) = \mathcal{E} (S_{f;\psi} S_{f;\psi}^T) + \mathcal{E} (S_{g;\psi} S_{g;\psi}^T) - \mathcal{E} (S_{g;\psi} S_{g;\tau}^T) (\mathcal{E} [S_{g;\tau} S_{g;\tau}^T])^{-1} \mathcal{E} (S_{g;\tau} S_{g;\psi}^T).$$

In the right hand side of this expression, according to (13), the last two terms give the F-information in  $g$ , and  $\mathcal{E} (S_{f;\psi} S_{f;\psi}^T) = I_f(\psi; \omega)$ , the F-information in  $f$ , because  $S_{f;\tau} = 0$ .

This results in:

$$I_p(\psi; \omega) = I_f(\psi; \omega) + I_g(\psi; \omega). \square$$

b. From the definitions and a property of conditional expectations follows:

$$I_f(\psi; \omega) = \mathcal{E}_X [S_{f;\psi} S_{f;\psi}^T] = \mathcal{E}_T [\mathcal{E}_{X|T} (S_{f;\psi} S_{f;\psi}^T | T)] = \mathcal{E} I_f(\psi | T), \text{ which gives:}$$

$$I_p(\psi; \omega) = \mathcal{E} I_f(\psi | T) + I_g(\psi; \omega). \square$$

## Chapter 3

### Loss of information in estimating item parameters in incomplete designs<sup>1</sup>

---

<sup>1</sup>The work in this chapter was done in cooperation with N.D. Verhelst. The chapter will be published as a Measurement and Research Department Reports 2004-3 Arnhem: Cito and will be submitted for publication in Psychometrika.

## **Abstract**

In this chapter, the efficiency of conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation of the item parameters of the Rasch model in incomplete designs is studied. The use of the concept of F-information (Eggen, 2000) is generalized to incomplete testing designs. The standardized determinant of the F-information matrix is used for a scalar measure of information in a set of item parameters. In this paper, the relation between the normalization of the Rasch model and this determinant is clarified. It is shown that in comparing estimation methods with the defined information efficiency is independent of the chosen normalization.

In examples, information comparisons are conducted. It is found that for both CML and MML some information is lost in all incomplete designs compared to complete designs. A general trend is that with increasing test booklet length the efficiency of an incomplete to a complete design and also the efficiency of CML compared to MML is increasing. The main differences between CML and MML is seen in relation to the length of the test booklet. It will be demonstrated that with very small booklets, there is a substantial loss in information (about 35%) with CML estimation, while this loss is only about 10% in MML estimation. However, with increasing test length, the differences between CML and MML quickly disappear.

### **3.1 Introduction**

In the Rasch model, two methods are commonly used to consistently estimate the item parameters: conditional maximum likelihood (CML) and marginal maximum likelihood (MML). In Eggen (2000)<sup>2</sup>, the loss of information resulting from the use of CML estimation was studied. The concept of F-information was shown to be useful in quantifying this loss. It was shown that CML estimation normally involves some loss of information with respect to the item parameters compared to using the full likelihood (JML), but the loss is small. Also a small loss was found in comparing the information in CML with MML estimation. The reported efficiencies of CML compared to JML and to MML are larger than 92%.

The results in Eggen (2000) only concern complete testing designs. In the present study the loss of information in incomplete designs will be treated. First, a short review of the concept of F-information and the expressions for CML and MML in the Rasch model will be given. For comparing information matrices and computing the information efficiency, the determinant criterion (see Chapter 2) was used. It will be shown that this criterion has some unique properties. Next, the generalization of the use of the concepts in incomplete designs in general and for the efficiency comparison in particular will be given. Some examples will be used to illustrate the loss of information and the efficiency of CML versus MML in incomplete designs with the Rasch model. A distinction will be made between the possible loss due only to the incompleteness of the design, and the loss due to the design and the estimation method combined.

### **3.2 F-information and the Rasch model**

In Eggen (2000), the F-information concept was used to study the loss of information in models with two (vector) parameters where there is interest in

---

<sup>2</sup>Eggen (2000) was reprinted in chapter 2 of this dissertation, but with some major revisions. If in this chapter there is a reference to chapter 2 this is to a revised part of the paper, otherwise the reference will be Eggen (2000).

only one of the two parameters and in which, by conditioning or marginalizing, one parameter is in a sense ignored in the inference. The F-information is defined as generalized Fisher information and it expresses the information in a two-parameter distribution with respect to one of the two parameters. If  $p(x; \omega)$  is the two-parameter distribution with  $\omega^T = (\psi^T, \tau^T)$  and  $S_{p; \omega}(X) = \partial \ln p(X; \omega) / \partial \omega$  is the efficient score statistic, then the Fisher information matrix is given by

$$I_p(\omega) = \mathcal{E}[S_{p; \omega} \cdot S_{p; \omega}^T] = \mathcal{E} \begin{bmatrix} S_{p; \psi} S_{p; \psi}^T & S_{p; \psi} S_{p; \tau}^T \\ S_{p; \tau} S_{p; \psi}^T & S_{p; \tau} S_{p; \tau}^T \end{bmatrix} = \begin{bmatrix} I_p^{\psi\psi} & I_p^{\psi\tau} \\ I_p^{\tau\psi} & I_p^{\tau\tau} \end{bmatrix}. \quad (1)$$

The F-information in  $p$  with respect to  $\psi$  is then given by

$$I_p(\psi; \omega) = \mathcal{E} S_{p; \psi} S_{p; \psi}^T - \mathcal{E} S_{p; \psi} S_{p; \tau}^T \cdot \mathcal{E} (S_{p; \tau} S_{p; \tau}^T)^{-1} \cdot \mathcal{E} S_{p; \tau} S_{p; \psi}^T. \quad (2)$$

Next consider the (vector) statistic  $T = T(X)$  with distribution  $g(t; \omega)$ , and write  $p(x; \omega)$  as the product of  $g(t; \omega)$  and the conditional distribution of  $x$  given  $t : f(x|t; \omega)$ . In Eggen (2000), it was shown that in two-parameter models which can be decomposed as  $p(x; \omega) = f(x | t; \psi) \cdot g(t; \omega)$ , the F-information with respect to  $\psi$  in  $p(x; \omega)$  is the sum of the F-information in the conditional model  $f(x|t; \psi)$ , which only depends on  $\psi$ , and the F-information in the model  $g(t; \omega)$ :  $I_p(\psi; \omega) = I_f(\psi; \omega) + I_g(\psi; \omega)$ . It is readily understood that if one uses only the conditional model for inference on  $\psi$  instead of the full model, there is no loss of information if  $I_g(\psi; \omega) = \mathbf{0}$ . Bhapkar (1989) has shown that a sufficient condition for this is that the distribution of  $T$  is at least weakly ancillary with respect to  $\psi$ . In the event the distribution is completely independent of  $\psi$ , the ancillary case, of course also no information is lost.

In the Rasch model for dichotomously scored items, the probability of getting the answer pattern  $x$ , with responses  $X_{vi} = x_{vi}$  (0 or 1) of persons  $v = 1, \dots, n$  on items  $i = 1, \dots, k$ , is given by

$$p(x; \beta, \theta) = \frac{\exp\{-\sum_i (\sum_v x_{vi})\beta_i + \sum_v (\sum_i x_{vi})\theta_v\}}{\prod_i \prod_v \{1 + \exp(\theta_v - \beta_i)\}}. \quad (3)$$

In (3),  $\beta^T = (\beta_1, \dots, \beta_k)$  is the item parameter and  $\theta^T = (\theta_1, \dots, \theta_n)$  the person or ability parameter.  $T_v = \sum_i X_{vi}$ , the sum score of a person is sufficient for  $\theta_v$ , for  $v = 1, \dots, n$  and the distribution of  $T$  is given by

$$g(t; \beta, \theta) = \prod_{v=1}^n g(t_v; \beta, \theta_v) = \prod_{v=1}^n \frac{\exp(\theta_v t_v) \cdot \gamma_{t_v}(\beta)}{\prod_{i=1}^k \{1 + \exp(\theta_v - \beta_i)\}}, \quad (4)$$

in which

$$\gamma_t(\beta) = \sum_{\sum y_i = t} \exp(-\sum_{i=1}^k \beta_i y_i) \quad (5)$$

are the so-called basic symmetric functions of order  $t$ . In (5), the summation runs across all answer patterns  $(y_1, \dots, y_k)$ ,  $y_i \in \{0, 1\}$  for which  $\sum_{i=1}^k y_i = t$ . It is easily checked that the conditional distribution of the answer patterns  $x$  given the scores  $t$  is only dependent on the item parameter:  $p(x; \beta, \theta) = f(x|t; \beta) g(t; \beta, \theta)$ ; in CML estimation of the item parameters only the conditional distribution is used.

Weak ancillarity is the key condition for losing no information in estimating the item parameters with CML instead of using the full model. When the model belongs to the exponential family, which the Rasch model (3) clearly does, the fulfillment of this condition is readily checked. (See Theorem 3 of Chapter 2 of this dissertation and Bhapkar (1989) for a proof of the theorem.)

$T$  is weakly ancillary for  $\beta$  if and only if there exist functions of  $\beta$  only and independent of the data,  $w_j(\beta)$ ,  $j = 1, \dots, k$ , and  $v(\beta)$ , such that:

$$\frac{\partial \ln \gamma_t(\beta)}{\partial \beta_j} = w_j(\beta) \cdot t + v(\beta), \text{ for } j = 1, \dots, k \quad (6)$$

for all  $t$ .



In the Rasch model, this partial derivative is given by (for  $j = 1, \dots, k$ ):

$$\frac{\partial \ln \gamma_t(\beta)}{\partial \beta_j} = - \frac{e^{-\beta_j} \cdot \gamma_{t-1}^{(j)}}{\gamma_t(\beta)}, \quad (7)$$

in which

$$\gamma_{t-1}^{(j)} := \partial \gamma_t(\beta) / \partial e^{-\beta_j}, \text{ for } j = 1, \dots, k. \quad (8)$$

It is well known (Molenaar, 1995) that expression (7) is equal to the conditional probability of answering item  $j$  correctly given the score  $t$ :  $P(X_j = 1 | t)$ . In general if  $t \neq 0$ , rewrite (7) as

$$\frac{\partial \ln \gamma_t(\beta)}{\partial \beta_j} = - \frac{e^{-\beta_j} \cdot \gamma_{t-1}^{(j)}}{\gamma_t(\beta)} \cdot \frac{t}{t} = - \frac{e^{-\beta_j} \cdot \gamma_{t-1}^{(j)}}{\sum_{i=1}^k e^{-\beta_i} \cdot \gamma_{t-1}^{(i)}} \cdot t, \quad (9)$$

in which the recursive formula  $\gamma_t(\beta) \cdot t = \sum_{i=1}^k e^{-\beta_i} \cdot \gamma_{t-1}^{(i)}$  (Fischer, 1974) is used. Then in expression (6),  $v(\beta) = 0$ , and  $w_j(\beta) = -e^{-\beta_j} \cdot \gamma_{t-1}^{(j)} / \sum_i e^{-\beta_i} \cdot \gamma_{t-1}^{(i)}$ , which is not only a function of the item parameters  $\beta$ , but also dependent on the data,  $t$ . Therefore,  $T$  is in general not weakly ancillary for  $\beta$ .

There is, however, an interesting case which yields weak ancillarity. This is the case if all item parameters are equal:  $\beta_j = \beta, j = 1, \dots, k$ . Then (9) simplifies to:

$$\frac{\partial \ln \gamma_t(\beta)}{\partial \beta_j} = \frac{e^{-\beta} \cdot \gamma_{t-1}^{(j)}}{k \cdot e^{-\beta} \cdot \gamma_{t-1}^{(j)}} \cdot t = \frac{1}{k} \cdot t. \quad (10)$$

This means that, in this special case, there is no loss of information in estimating the item parameters if CML is used instead of the full model.

For the Rasch model, the expressions for the F-information with respect to the item parameter  $\beta$  are given in Eggen (2000). For instance, in the full Rasch model (3), evaluating expression (2) gives the F-information. The expressions in the cases of MML and CML estimation of the item parameters are given in full detail in Eggen (2000).

If MML estimation of the item parameters in the Rasch model is used, we have a two-parameter distribution  $p_m(x; \beta, \xi)$ , the first parameter being the item parameter  $\beta$  and the second,  $\xi$ , the parameters of the ability distribution. In this case,  $\omega^T = (\beta^T, \xi^T)$  and the F-information with respect to  $\beta$  is formally given by

$$I_{p_m}(\beta; \omega) = I_{p_m}^{\beta\beta^T} - I_{p_m}^{\beta\xi^T} [I_{p_m}^{\xi\xi^T}]^{-1} I_{p_m}^{\xi\beta^T}. \quad (11)$$

The F-information is expressed in terms of the Fisher information matrix (1). In this case,  $\psi = \beta$  and  $\tau^T = \xi^T$ . So,  $I_{p_m}^{\beta\beta^T}$  is then the item parameter submatrix of the Fisher information matrix.  $I_{p_m}^{\xi\xi^T}$  is the part of the ability distribution parameters. If we use a normal distribution, this matrix has four elements ( $\xi^T = (\mu, \sigma^2)$ ). And, finally,  $I_{p_m}^{\beta\xi^T}$  is the part in which each element includes a partial derivative of the likelihood with respect to both an item parameter and an ability distribution parameter.

In the case of CML estimation of the item parameters, the conditional distribution of the answer pattern given the sufficient statistic,  $f(x|t; \beta)$ , is only dependent on  $\beta$  and the F-information in this case is equal to the expected conditional Fisher information:

$$I_f(\beta; \omega) = \mathcal{E}[S_{f\beta} \cdot S_{f\beta}^T] = \mathcal{E} \left( \frac{\partial \ln f(x|t; \beta)}{\partial \beta} \cdot \frac{\partial \ln f(x|t; \beta)}{\partial \beta}^T \right) = \mathcal{E}_T I_f(\beta | T). \quad (12)$$

In which the latter expectation is taken with respect to the distribution of  $T$ , which is usually derived from the distribution of the ability  $\theta$ . If a normal ability distribution with mean  $\mu$  and variance  $\sigma^2$  is used, the exact expressions for computing the F-information for CML and MML estimation in the Rasch model are given in Eggen (2000).

It should be understood that the expressions for the F-information matrices in the Rasch model, given in (11) and (12), are only useful, if the model is properly normalized. It is well known that in the Rasch model not all  $k$  but only  $(k-1)$  item parameters are free. A commonly used normalization is to fix the value of one parameter to zero: e.g.  $\beta_1 = 0$ . In Chapter 2 of this dissertation this

normalization was also used and a scalar quantification of the F-information matrices was then given by their determinants. And, following Pukelsheim (1993), the information efficiency was used to compare the F-information in two models. For comparing MML and CML estimation of  $k$  items in the Rasch model, this is computed as

$$\text{INFEFF}(\beta; f: p_m) = \left( \frac{\det(\mathbf{I}_f^{*(i)}(\beta; \omega))}{\det(\mathbf{I}_{p_m}^{*(i)}(\beta; \omega))} \right)^{1/(k-1)}, \quad (13)$$

in which  $\mathbf{I}^{*(i)}$  denotes a F-information matrix after normalization on item  $i$ . In the next section, the relation between the normalization of the model and the information matrices will be treated in detail.

### 3.3 Normalization, information, and the determinant

In Eggen (2000), it was seen that the F-information matrices of CML and MML estimation of the item parameters in the Rasch model are double centered and therefore not of full rank. This is a consequence of the well-known indeterminacy in the model, which can be solved by imposing one linear restriction on the parameters (Molenaar, 1995). This normalization of the parameters always has an influence on the information matrix. In order to compare different estimation methods on the information, it seems to be reasonable to use the same normalization. In the sequel it will become clear that for comparing this condition can be relaxed somewhat.

In MML, a common normalization is to set the population mean  $\mu$  equal to some constant. But when CML is used, the population mean does not need to be defined and, consequently this normalization cannot be used for comparing CML and MML. Therefore, only normalizations on the item parameters will be considered.

Proper normalizations of the parameter space in the Rasch model can be characterized by the following equation:

$$d_0 + \sum_{i=1}^k d_i \beta_i = 0, \text{ with the restriction that } \sum_{i=1}^k d_i \neq 0. \quad (14)$$

Without loss of generality, it is assumed in the sequel that  $d_0 = 0$ .

The most commonly used normalizations are twofold. The first is that one of the item parameters is fixed to an arbitrary constant,

$$\beta_i = 0. \quad (15)$$

The other is that a linear restriction is put on the mean or the sum of the item parameters, for instance,  $\sum_i \beta_i = 0$ .

It is easily checked that both normalizations are special cases of (14) (with  $d_0 = 0$ ).

The normalization in (14) is equivalent with saying that there will be at least one  $i$  with  $d_i \neq 0$ , such that  $\beta_i$  can be written as

$$\beta_i = \sum_{j \neq i} c_j \beta_j, \quad (16)$$

with the restriction that  $\sum_{j \neq i}^k c_j \neq 1$  and with  $c_j = -d_j/d_i$ . Henceforth in this chapter, (16) will be used for the normalization.

### 3.3.1 The influence of the normalization on the information matrices

The F-information matrices of the item parameters in the Rasch model given in (11) for MML and in (12) for CML estimation are matrices in which the normalization is discarded. These  $(k \times k)$ -matrices are determined as if all  $k$  item parameters are free. These double-centered matrices will be called the non-normalized F-information matrices. If the model is properly normalized, it has only  $k-1$  free parameters, and the normalized F-information matrices will be  $(k-1) \times (k-1)$ . Next, we will express the normalized in the non-normalized F-information.

Although normalization by fixing one item parameter (15) is a special case of normalization (16), this case will be treated first because, in this case, the results are easy and straightforward. Furthermore, it will be shown that for the computation of the information efficiency in the more general case, the results

of this special case suffice.

### Normalization by fixing one item

A normalized F-information matrix is deduced from the non-normalized F-information by simply deleting the row and the column corresponding to the fixed item. For the determinant of such a matrix, the following lemma holds.

#### **Lemma 1**

Let  $A$  be a  $k \times k$  matrix which is double centered, that is,

$$A\mathbf{1} = \mathbf{0} \text{ and } \mathbf{1}^T A = \mathbf{0}^T \quad (17)$$

and  $A^{(i)}$ , a principal submatrix of  $A$ , which results if the  $i^{th}$  row and the  $i^{th}$  column of  $A$  are deleted. Then the determinants of all principal submatrices, or all the principal minors, are equal:

$$\det A^{(i)} = \det A^{(j)}, \quad i, j = 1, \dots, k. \quad (18)$$

The proof of the lemma is in the appendix of this chapter (p.96).  $\square$

If an F-information matrix after a normalization by fixing item  $i$  is denoted by  $I^{\star(i)}(\beta; \omega)$ , then it yields  $\det(I^{\star(i)}(\beta; \omega)) = \det(I^{\star(j)}(\beta; \omega))$ , for  $i, j = 1, \dots, k$ . This is true for the F-information matrices in the MML (11) and in the CML (12) case. Consequently, not only the determinant of the F-information, but also the information efficiency (13), used in comparing MML and CML, is independent of the item which is fixed in the normalization:

$$\left( \frac{\det(I_f^{\star(i)}(\beta; \omega))}{\det(I_{p_m}^{\star(j)}(\beta; \omega))} \right)^{1/(k-1)} = \left( \frac{\det(I_f^{\star(l)}(\beta; \omega))}{\det(I_{p_m}^{\star(m)}(\beta; \omega))} \right)^{1/(k-1)}, \text{ for } i, j, l, m = 1, \dots, k. \quad (19)$$

### Normalization by a general linear restriction

If the normalization is established by putting a linear restriction on the item parameters as in (16), the non-normalized F-information matrices in (11) and (12) can also be expressed in the normalized F-information matrices. Denote by  $S_{p; \omega}^{(i)}$

the efficient score  $(k-1)$ -vector of distribution  $\mathbf{p}$  with respect to parameter  $\boldsymbol{\omega}$  if the normalization is on item  $i$  as in (16). (In the MML case  $\boldsymbol{\omega}^T = (\boldsymbol{\beta}^T, \boldsymbol{\xi}^T)$ ). The elements of this score vector have a simple relation with the elements of the non-normalized score vector  $\mathbf{S}_{\mathbf{p};\boldsymbol{\omega}}$ .

The partial derivatives, under the restriction (16), with respect to an item parameter  $\beta_j^*$  ( $j \neq i$ ) of the log likelihood (in both the MML and CML case) are expressed in the non-restricted derivatives as

$$\frac{\partial \ln L}{\partial \beta_j^*} = \sum_{l=1}^k \frac{\partial \ln L}{\partial \beta_l} \frac{\partial \beta_l}{\partial \beta_j^*} = \frac{\partial \ln L}{\partial \beta_j} + c_j \frac{\partial \ln L}{\partial \beta_i}, \text{ for } j = 1, 2, \dots, i-1, i+1, \dots, k. \quad (20)$$

Under the restriction (16) there are only  $(k-1)$  free item parameters  $\beta_j^*$  instead of  $k$ . In the sequel we will drop the  $*$  from the notation of the item parameters, but it should be understood that in all normalized information matrices only  $(k-1)$  item parameters are considered.

Without loss of generality it is assumed in the sequel that  $i = 1$  in (16) and (20).

If we define a  $k \times (k-1)$  matrix  $\mathbf{K}$  by

$$\mathbf{K} = \begin{bmatrix} \mathbf{c}^T \\ \mathbf{I}_{k-1} \end{bmatrix}, \text{ with } \mathbf{c}^T = (c_2, \dots, c_k), \quad (21)$$

then it is seen that, in the CML case,

$$\mathbf{S}_{f;\boldsymbol{\beta}}^{(i)T} = \mathbf{S}_{f;\boldsymbol{\beta}}^T \mathbf{K}. \quad (22)$$

We see that the normalized F-information matrix, denoted by  $\mathbf{I}_f^{(i)}(\boldsymbol{\beta}; \boldsymbol{\omega})$ , using (12) and (22), is related to the non-normalized F-information matrix as

$$\mathbf{I}_f^{(i)}(\boldsymbol{\beta}; \boldsymbol{\omega}) = \mathbf{K}^T \mathbf{I}_f(\boldsymbol{\beta}; \boldsymbol{\omega}) \mathbf{K}. \quad (23)$$

In the MML case with the normalization (16), the population parameters  $\boldsymbol{\xi}$  are free. Therefore we consider the  $(k+2) \times (k+1)$ -matrix  $\mathbf{K}^*$  :

$$K^* = \begin{bmatrix} K & \mathbf{0}_{(k \times 2)} \\ \mathbf{0}_{(2 \times k-1)}^T & I_2 \end{bmatrix} \quad (24)$$

and it is easily checked that

$$S_{p_m; \omega}^{(i)T} = S_{p_m; \omega}^T K^*. \quad (25)$$

With (25), it follows that, in the MML case, the relation between the normalized Fisher information matrix, to denote by  $I_{p_m}^{(i)}(\omega)$ , and the non-normalized is

$$I_{p_m}^{(i)}(\omega) = K^{*T} I_{p_m}(\omega) K^*. \quad (26)$$

And by using (11), it can be shown that the relation between the F-information matrix in the normalized case, denoted by  $I_{p_m}^{(i)}(\beta; \omega)$ , and the non-normalized is given by

$$I_{p_m}^{(i)}(\beta; \omega) = K^T I_{p_m}(\beta; \omega) K. \quad (27)$$

A second lemma on the determinant of double-centered matrices with a special structure is given below.

### **Lemma 2**

Let  $A^*$  be a symmetric  $k \times k$  matrix partitioned as follows:

$$A^* = \begin{bmatrix} \alpha & a^T \\ a & A \end{bmatrix} \quad (28)$$

$$\text{with } a = -A \mathbf{1} \text{ and } \alpha = -\mathbf{1}^T a = \mathbf{1}^T A \mathbf{1}. \quad (29)$$

So  $A$  is a  $(k-1) \times (k-1)$  matrix.

Next, consider a  $(k-1)$  vector  $c^T = (c_2, \dots, c_k)$  and a  $k \times (k-1)$ -matrix  $K$  with the following structure:

$$K = \begin{bmatrix} c^T \\ I_{k-1} \end{bmatrix}. \quad (30)$$

Then, for any double-centered matrix  $A^*$  with the structure given in (28) and (29) and for any matrix  $K$  as given in (30), it holds that

$$\det(K^T A^* K) = g(c) \det A, \quad (31)$$

with  $g(c) = (\det(I - c\mathbf{1}^T))^2$  a function of  $c$ , which is independent of  $A^*$ .

The proof of the lemma is in the appendix of this chapter (p.96).  $\square$

It will be clear that the normalized F-information matrices for CML (23) and for MML (27) have exactly the structure of the matrix presented in lemma 2 and thus yields (31). For the determinants of the F-information matrices, we then have that, for CML,

$$\det I_f^{(i)}(\beta; \omega) = \det(K^T I_f(\beta; \omega) K) = g(c) \cdot \det I_f^{*(i)}(\beta; \omega), \quad (32)$$

in which  $I_f^{*(i)}(\beta; \omega)$  is constructed as before from the non-normalized CML F-information matrix by dropping the row and the column corresponding to the item  $i$  which is used for normalization.

For MML we have similarly:

$$\det I_{p_m}^{(i)}(\beta; \omega) = \det(K^T I_{p_m}(\beta; \omega) K) = g(c) \cdot \det I_{p_m}^{*(i)}(\beta; \omega). \quad (33)$$

Note that both (32) and (33) yield for any  $i = 1, \dots, k$  used in the normalization. It should be understood that the F-information matrices change when the normalization is on a different item, but, in the determinants, these differences will only be reflected in different  $g(c)$ . This is because (see lemma 1) the principal minors of the non-normalized F-information matrices are independent of the item  $i$ .

If we compare CML with MML estimation, using the same normalization, this has the consequence that the computed efficiency is completely independent of



the normalization chosen. Thus, for  $i = 1, \dots, k$ :

$$\left( \frac{\det(I_f^{(i)}(\beta; \omega))}{\det(I_{p_m}^{(i)}(\beta; \omega))} \right)^{1/(k-1)} = \left( \frac{g(c) \cdot \det(I_f^{*(i)}(\beta; \omega))}{g(c) \cdot \det(I_{p_m}^{*(i)}(\beta; \omega))} \right)^{1/(k-1)} = \left( \frac{\det(I_f^{*(i)}(\beta; \omega))}{\det(I_{p_m}^{*(i)}(\beta; \omega))} \right)^{1/(k-1)} \quad (34)$$

It can be noted, that the strict condition to use the same normalization in computing the efficiency can be relaxed somewhat. It suffices to demand that the normalization should result in the same value of  $g(c)$ .

This result, combined with (19), implies that any efficiency comparison on the estimation of the item parameters with the determinants of the F-information matrices is independent of the normalization of the Rasch model.

In Eggen (2000), the trace function was used to compare the F-information matrices. The trace is very easily computed, but is not as useful as the determinant for comparing. If we look at the relation of the normalized and the non-normalized F-information, using the general notation of Lemma 2, then it is easily shown that

$$(K^T A K)_{ii} = A_{ii} + 2c_i a_i + c_i^2 \alpha \quad (35)$$

with  $a_i$  and  $c_i$  the  $i^{th}$  element of  $a$  and  $c$  respectively. Because of (35)

$$\text{tr}(K^T A K) = \text{tr}(A) + 2c^T a + \alpha c^T c. \quad (36)$$

It can be seen that there is no simple relation between the traces of the matrices. Consequently the comparison of F-information matrices with the trace function will always be dependent on the normalization chosen.

### 3.4 F-information in incomplete designs

In incomplete testing designs, not all items are administered to all persons. Define the design variable  $d_{vj} = 1$  if item  $j$ ,  $j = 1, \dots, k$ , was administered to person  $v$ ,  $v = 1, \dots, n$  and  $d_{vj} = 0$ , otherwise. The design vector of a person is denoted by  $d_v^T = (d_{v1}, \dots, d_{vk})$  and the complete design vector by  $d^T = (d_1^T, \dots, d_n^T)$ . Furthermore, assume that the response variable  $X_{vj}$  takes an arbitrary constant value  $c$  if an item  $j$  is not answered by person  $v$ :  $X_{vj} = c$  if  $d_{vj} = 0$ ; if  $d_{vj} = 1$  the values of the response variable are, as in the complete data case, equal to 1 or 0.

In the Rasch model, the probability of getting answer pattern  $x$  given the design vector  $d$  is given by

$$p(x|d; \beta, \theta) = \prod_{v=1}^n \prod_{i=1}^k p(x_{vi}|d_{vi}; \beta_i, \theta_v) = \prod_{v=1}^n \prod_{i=1}^k \frac{\exp[(\theta_v - \beta_i)x_{vi}d_{vi}]}{\{1 + \exp(\theta_v - \beta_i)\}^{d_{vi}}} \quad (37)$$

In general, one could consider the design variable to be random with distribution  $p(d; \phi)$ . Throughout this chapter it will be assumed that this distribution is independent of the parameters  $\beta$  and  $\theta$ . So, for inference on the parameters  $\beta$  and  $\theta$ , it is justified to consider only the conditional distribution of the answer pattern given the design, as specified in (37) instead of the simultaneous distribution of  $(x, d)$ . This can also be understood in terms of the loss of F-information. The following decomposition is true:

$$p(x, d; \beta, \theta, \phi) = p(x|d; \beta, \theta) p(d; \phi) \quad (38)$$

The simultaneous distribution of  $(x, d)$  is a two-parameter distribution: the first parameter is  $(\beta, \theta)$  and the second is  $\phi$ . Because the design distribution is assumed to be independent of the first parameter, the ancillary case, no information is lost if only the conditional distribution,  $p(x|d; \beta, \theta)$ , is considered in the inference on  $(\beta, \theta)$ . For more details on stochastic designs, see Rubin (1976) and also Chapter 4 of this dissertation.

The generalization to incomplete designs of CML and MML estimation of the item parameters is well known (Molenaar, 1995). In CML, the conditional distribution of the answer pattern given the scores,  $T_v = \sum_i d_{vi} X_{vi}$  for  $v = 1, \dots, n$ , and the design used is:

$$f(x|t, d; \beta) = \prod_{v=1}^n f(x_v|t_v, d_v; \beta) = \prod_{v=1}^n \frac{\exp[-\sum_{i=1}^k d_{vi} x_{vi} \beta_i]}{\gamma_{t_v}(\beta \odot d_v)}, \quad (39)$$

in which

$$\gamma(\beta \odot d_v) = \sum_{\sum_i d_{vi} y_i = t_v} \exp \left( - \sum_{i=1}^k y_i d_{vi} \beta_i \right) \quad (40)$$

are the basic symmetric functions as in complete designs (5). However, in (40), they are defined for the item parameter vector multiplied directly with the design vector:  $\beta \odot d_v = (\beta_1 d_{v1}, \dots, \beta_i d_{vi}, \dots, \beta_k d_{vk})^T$ , which means that only those items will be taken into account which are administered to a person. Furthermore, the summation in (40) runs across all answer patterns  $(y_1, \dots, y_k), y_i d_{vi} \in \{0, 1\}$  for which  $\sum_{i=1}^k y_i d_{vi} = t_v$ .

If the ability distribution is denoted by  $h(\theta; \xi)$  then the marginal distribution, used in MML estimation is given by:

$$\begin{aligned} p_m(x|d; \beta, \xi) &= \prod_{v=1}^n \int_{-\infty}^{\infty} \prod_{i=1}^k p(x_{vi} | \theta, d_{vi}; \beta_i) h(\theta; \xi) d\theta \\ &= \prod_{v=1}^n \int_{-\infty}^{\infty} \prod_{i=1}^k \frac{\exp[(\theta - \beta_i) x_{vi} d_{vi}]}{\{1 + \exp(\theta - \beta_i)\}^{d_{vi}}} \cdot h(\theta; \xi) d\theta \end{aligned} \quad (41)$$

The F-information with respect to the item parameters in CML as well as in MML estimation is defined as expected information. In efficiency comparisons in complete testing designs, it suffices therefore to derive this information for one randomly selected person from the ability distribution. There is no need to consider more persons because each has an independent equal expected contribution to the F-information. This is not true in incomplete designs because there is no contribution to the information on  $\beta_i$  if  $d_{vi} = 0$ . Therefore, the total information in the whole sample of persons  $v = 1, \dots, n$ , will be considered, i.e., the design must be taken in account. An elegant and easy way to avoid complicated formulae which contain the design indicators  $d_{vi}$ , is to consider the contribution of every person  $v$  with design vector  $d_v$  separately, as displayed in Figure 3.1.

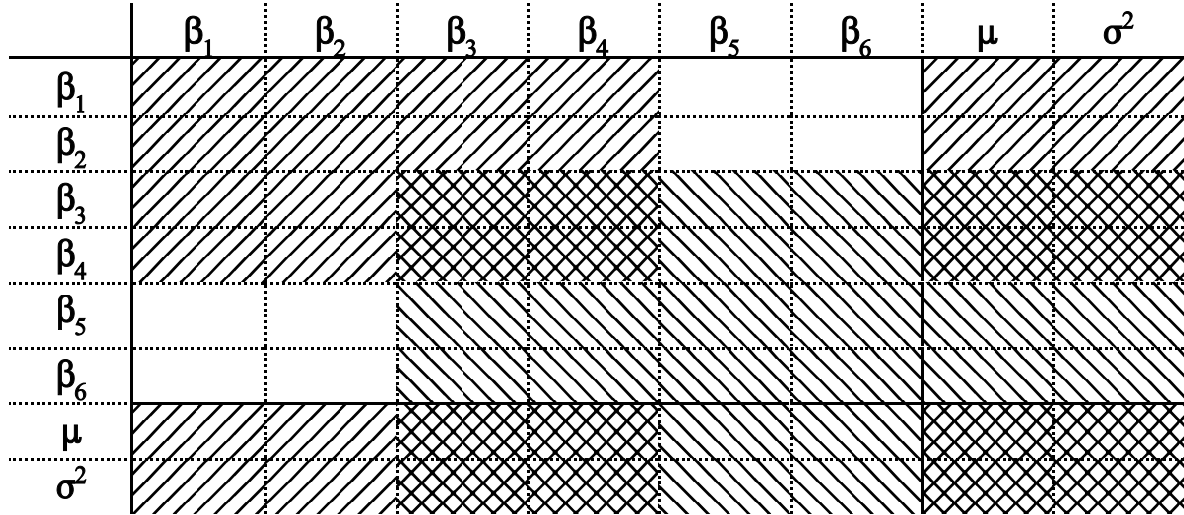




Figure 3.1. Schematic representation of the elements of the information matrices in incomplete designs by the example  $\mathbf{d}_1^T = (1, 1, 1, 1, 0, 0)$  , and  $\mathbf{d}_2^T = (0, 0, 1, 1, 1, 1)$  .

With design vector  $\mathbf{d}_1$ , only the items 1 to 4 are administered, and for such an administration the information matrix can be computed as in a complete design with these four items. It will be clear that with such a test no information is collected with respect to the items 5 and 6. Using design vector  $\mathbf{d}_2$  (independently to another person randomly drawn from the same ability distribution), and applying the same procedure we now collect information on the parameters of the items 3 to 6 and on the two population parameters. Because of the independence, the information from both observations can be added, leading to the situation as depicted in Figure 3.1: for some cells (doubly shaded) the two observations have a contribution, for others (singly shaded) only one observation contributes, and for some of the cells, there is no contribution at all. For example, there is no information available on the items 1 and 5 jointly, because this pair has not been observed jointly.

Notice that the contribution of several design vectors to the same cell is not equal. Consider, for example, the cell on the main diagonal with respect to the

parameter  $\mu$ : the amount of information on this parameter depends on the number of item parameters and their specific value for each design vector  $d_v$ .

### 3.5 The information comparison in incomplete designs

In complete testing designs, only one randomly selected person from the ability distribution was needed to compare the F-information matrices for two models, and the comparison of two models was the only sensible comparison to make. With incomplete designs, however, efficiency comparisons can be made along several lines: one could compare the efficiency of two models, given that the data will be collected in a given design, but one could also compare the efficiency of two different designs under the same model.

A good efficiency measure for the first problem is (13), where the information matrix  $I^{*(i)}$  can represent the total F-information from a sample of size  $n$  with data collected in some (incomplete) test design. As long as  $n$  and the design are the same under the two models, the results will be unaffected by the sample size. This can easily be understood from a simple example. Suppose we collect data using a complete design and the sample size  $n = 2$ . The elements of the total information matrix under either model will be twice the value of the elements in the matrix for a single person and consequently the determinants of both total F-information matrices will be  $2^{(k-1)}$  times the value of the determinants of the single observation, but the factor  $2^{(k-1)}$  will cancel in numerator and denominator of the right hand side of (13). Of course, the efficiency of CML compared to MML may vary according to the design.

Before presenting a more general efficiency measure, we first discuss the practical implication of the exponent  $1/(k-1)$  in the efficiency measure (13). The  $(k-1)$ -th root of the determinant of the normalized F-information matrix under some model, where the information matrix is computed using a single person, can be understood as the total amount of information per person. So we could define

$$m_*(I_{1,f}) = [\det(I_f(\beta; \omega))]^{1/(k-1)} \quad (42)$$

where  $f$  refers to the model,  $k$  is the number of items,  $*$  refers to the

normalization used and the '1' in subscript refers to the number of observations used in deriving the information matrix. From now, in the notation of the information matrices the indices referring to the normalization are dropped. Then it is easy to show that

$$m_*(I_{t,f}) = t \cdot m_*(I_{1,f}) \quad (43)$$

Since the information matrix is additive in the number of persons, the entries in the non-normalized as well as the normalized information matrices based on  $t$  independent observations are  $t$  times the corresponding values with a single person. The normalized information matrix has  $(k-1)$  rows and columns and therefore the determinant of the normalized information matrix based on  $t$  persons will be  $t^{(k-1)}$  times the determinant of the matrix based on a single person. By taking the  $(k-1)$ -th root, the assertion follows.

If we want to compare two designs under the same model, there is no simple and unequivocal measure of efficiency, as can be seen from the following simple example. Suppose we want to determine the efficiency of the design used in Figure 3.1 compared to a complete design. To apply the incomplete design, one needs at least a sample size of 2, while the complete design can be used with a single observation. If one uses these numbers, the incomplete design may appear as the most efficient one, but one could argue that the comparison is not fair because of unequal sample sizes. If on the other hand, one uses equal sample sizes, the complete design will appear the most efficient because in the incomplete design either of the two persons has answered only to a subset of the  $k$  items. So, to make a fair comparison, it is not sufficient to consider only the differences in sample sizes.

It seems reasonable, therefore, to develop an efficiency measure which takes in some respect the cost of the design implementation into account, by expressing the information measure relative to the cost of implementation of the design. The measure of efficiency is then the ratio of these two relative information measures. So the measure proposed in (13) can then be generalized to

$$\text{INFEFF}(\beta; f(D_1); p_m(D_2), C) = \frac{C(D_2)}{C(D_1)} \left( \frac{\det(I_{f(D_1)}(\beta; \omega))}{\det(I_{p_m(D_2)}(\beta; \omega))} \right)^{1/(k-1)}. \quad (44)$$

Where  $C(D)$  represents the cost of the data collection according to the design  $D$  and  $I_{f(D)}(\beta; \omega)$  represents the F-information matrix under model  $f$  using design  $D$ , under a proper normalization which is common to both models.

Formula (13) is easily seen as a special case, where the two models differ but the designs are the same, so that their cost ratio equals one.

In the next section, examples will be studied for another special case of (44), where the designs differ, but the models are the same, either CML or MML, so that we obtain a measure of the relative efficiency of two designs. An interesting question in this respect is whether and to what extent this measure will vary across different models.

An important issue is the definition of the cost function  $C(D)$ , which can range from a simple model with a fixed unit cost per item answer, to more realistic functions with different costs per test taker and per answer within a test taker up to complicated functions which take the risk of errors in the implementation of the design into account.

Within the present chapter we will confine ourselves to two cost functions: a cost function with a unit cost per observed answer, and another one with a unit cost per test taker. We illustrate both cases using a simple example, where the design of Figure 3.1 is compared to the complete design.

Consider the first cost function ( $C_1$ ) with a single response of a person to an item as unit cost. Suppose the information measure in the complete design is based on  $n_c$  test takers, and the information in the incomplete design on  $n_i$  test takers. We will assume for simplicity that  $n_i$  is an integer multiple of the number of different design vectors used in the incomplete design. Then it will be clear that (see also Figure 3.1)  $C_1(D_c) = 6 \cdot n_c$ , while  $C_1(D_i) = 4 \cdot n_i$ .

Application of (44) yields a comparison of both designs in terms of amount of information per single response. Since both cost functions are multiples of the sample sizes, and in view of (43) it will be clear that the efficiency measure (44)

in combination with cost function  $C_1$  is invariant under the choice of any (positive) sample sizes, either in the complete as in the incomplete design. Although such a measure may not be directly useful in planning data collections, it gives a kind of intrinsic comparison between designs. In fact, it can give a sensible answer to the question whether a single response contributes the same amount of information to the estimation of the item parameters in an incomplete design compared to another incomplete design or to a complete design.

The second cost function  $C_2$ , which is equal or proportional to the sample sizes, yields a more practical result. It is useful for comparing the efficiency of two designs with the same model. The efficiency measure compares the amount of information per test taker in two different designs. This means that the cost functions of the design in Figure 3.1 and the complete design are respectively:  $C_2(D_i) = n_i$  and  $C_2(D_c) = n_c$ . Suppose that for the example of Figure 3.1 (and a given set of parameter values) we find  $\text{INFEFF}(\beta; f(D_i):f(D_c), C_2) = 0.8$ . This means that the amount of information delivered per test taker in the incomplete design is 80% of the information per test taker in the complete design, or, conversely, if we would plan a data collection with 1000 test takers in the complete design, we would need  $1000/0.8 = 1250$  test takers in the incomplete design to collect the same amount of information with the incomplete design.

As a final note, it should be stressed that equating the total amount of information in two different designs by manipulating the sample sizes, does not mean that these two designs are therefore equivalent in all possible respects. This can easily be seen from Figure 3.1: whatever the sample size in the incomplete design used there, there will be always eight zero cells (the cells without any shading). It is beyond the scope of the present chapter, however, to predict in a detailed way all possible implications of such and similar patterns in the information matrix.

### 3.6 Examples of comparing the efficiency in incomplete designs

The efficiency comparisons are conducted with two item banks, each with 18 items. The item bank size of 18 was chosen because at that point the difference



in efficiency of CML compared to MML in complete designs (Eggen, 2000) is negligibly. In all examples, a normal ability distribution with mean 0 and variance 1 was used. Two different sets of item parameters were considered:

1. all items equal to 0:  $\beta_i = 0, i = 1, \dots, 18$
2. symmetric around 0, ranging from -2.25 to 2.25 in steps of 0.25, omitting 0 itself. This will be denoted by  $\beta_i = -2.25 (0.25) 2.25, i = 1, \dots, 18$ .

With the first set of items, there is, as was seen before, no loss of information in using CML compared to MML, and the sole loss due to the incompleteness can be studied. In the second situation, the extra loss resulting from the use of CML can be considered.

### **3.6.1 The designs**

The incomplete designs considered are in common use and are all so-called anchor-item designs. In order to be able to estimate the item parameters from the resulting data, there is an overlap between the tests or booklets administered. A booklet is a number of items which are administered together to a person. All the designs considered share the properties that every booklet is the same length and that every item has the same number of observations. Furthermore, it is assumed that the booklets are randomly assigned to the students. Three different types of designs are distinguished:

1. Item-interlaced anchoring design (Vale, 1986). In these designs there are as many booklets as there are items. The first booklet begins with the first item and contains sequentially all items until it has its established test length. The second booklet starts with the second item. The final booklet contains the last item and one or more of the first items. These designs can have tests or booklets with of 2, 3, or  $k-1$  items. The limiting case with  $k$  items gives a complete design. In the sequel, the following notation will be used for these designs: II;#i;#b, in which II is the abbreviation of the type of design, #i is the length of a booklet in the design and #b is the number of booklets in the design. In Figure 3.2 two examples are given in case the total number of items is  $k=18$ .

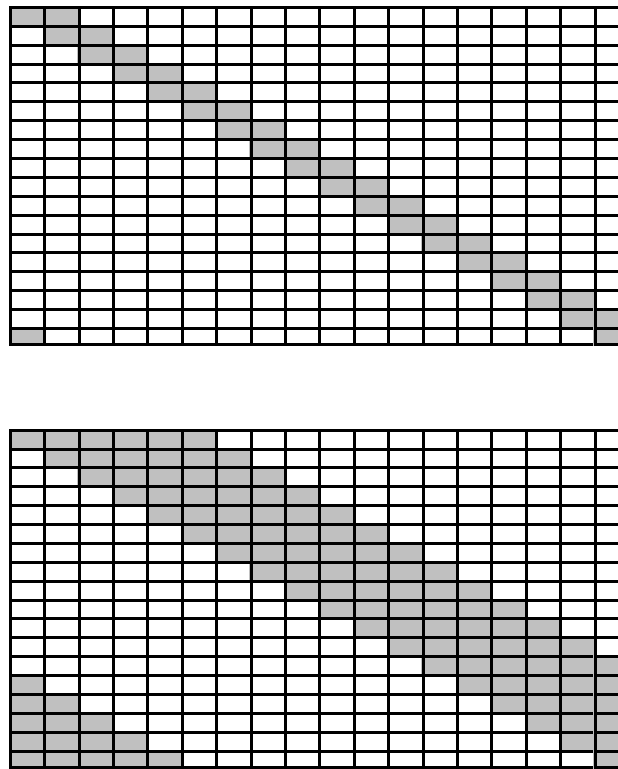


Figure 3.2. Item-interlaced anchoring designs II;2;18 and II;6;18

In design II;2;18: the lengths of the booklets are 2: the first booklet contains item number 1 and 2, the second item number 2 and 3, etcetera, until the eighteenth booklet with item number 18 and 1. In II;6;18 the test lengths are 6: booklet 1 has the items 1, 2, 3, 4, 5, and 6; booklet 2 the items 2, 3, 4, 5, 6, and 7, and so on.

2. Block-interlaced anchoring design. In these designs, a booklet contains two blocks of items. These two blocks are interlaced as in an item-interlaced anchoring design with booklets with two blocks. Each block can contain more than one item, but the number of items per block are equal. The general notation is BI;#i;#b, with #i the total number of items in the two blocks.

In the case of 18 items, 4 of these designs can be established. The block size is 1, 2, 3, or 6 items. These designs then have respectively 18 booklets with 2 items, 9 booklets with 4 items, 6 booklets with 6 items, and 3 booklets with 12 items. The block-interlaced design with 6 booklets, each having two

blocks of 3 items is given in Figure 3.3.

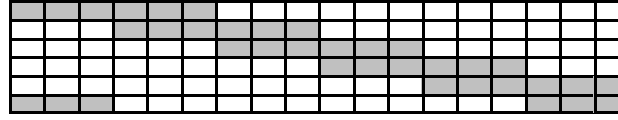


Figure 3.3. Block-interlaced anchoring design  $BI;6;6$

Remark that the item interlaced design  $II;2;18$  in Figure 3.2 is the same as the block interlaced design  $BI;2;18$  with blocksize 1.

3. Balanced block designs. In balanced block designs, not only all the blocks of items have an equal number of observations. Also, the pairs of blocks of items have an equal number of observations. The notation we will use for that is  $BB;(\#bl.bls),\#b$ , in which  $\#bl.bls$  is the number of blocks in a booklet multiplied by the size of the blocks. (This is of course the number of items in a booklet). With 18 items in total, there are several balanced block designs possible (Cochran & Cox, 1957). Only designs with no more booklets than items are considered. Two examples are given in Figure 3.4.

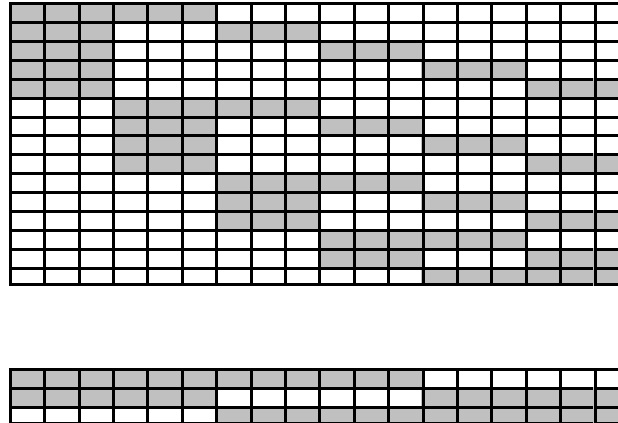


Figure 3.4 Balanced block designs  $BB;(2.3);15$  and  $BB;(2.6);3$

It can be seen that the  $BB;(2.6);3$  is the same as the block interlaced design  $BI;12;3$ . The balanced block designs with 18 items and no more than 18 booklets considered next are given in Table 3.1.

Table 3.1 Balanced block designs with 18 items and no more than 18 booklets

code	total number of blocks	block size	number of blocks per booklet	length of booklet	number of booklets
BB;(17.1);18	18	1	17	17	18
BB;(8.2);9	9	2	8	16	9
BB;(5.3);6	6	3	5	15	6
BB;(2.6);3	3	6	2	12	3
BB;(5.2);18	9	2	5	10	18
BB;(3.3);10	6	3	3	9	10
BB;(4.2);18	9	2	4	8	18
BB;(2.3);15	6	3	2	6	15

### 3.6.2 Results for an observed response as unit of cost

Using the cost function  $C_1$ , see section 3.5, comparisons between different estimation methods and different designs are possible. It is noted that with this cost function all the results are given per observed response. In the Tables 3.2 to 3.7, the information per observed response and the information efficiency will be given. The entries in the columns CML-det and MML-det are  $\text{CML-det} = 10^3 \cdot [\det(\mathbf{I}_{f(D)})]^{1/(k-1)} / C_1(D)$  and  $\text{MML-det} = 10^3 \cdot [\det(\mathbf{I}_{p_m(D)})]^{1/(k-1)} / C_1(D)$  respectively.

In computing the determinants and the efficiencies (44) the normalization was established by fixing the first item.

#### Results for item-interlaced anchoring designs

For 18 items with  $\beta_i = 0$  in the Rasch model the standardized determinants of the F-information matrices per observation for CML and MML are given, followed by the information efficiency of CML compared to MML. For reference, the results for a complete design are also reported in Table 3.2. At each entry in Table 3.2 also the efficiency of using the design compared to the complete design, using the same estimation method, is given between brackets.

Table 3.2 Information comparison for item-interlaced anchoring designs;  $\beta_i = 0, i=1, \dots, 18$ .

design	CML-det (% of complete)	MML-det (% of complete)	INFEFF( $\beta; f(D):p_m(D), C_1$ )
complete	9.68	9.68	1
II;2;18	3.40 (0.351)	8.28 (0.855)	0.411
II;3;18	5.49 (0.567)	8.49 (0.877)	0.646
II;4;18	6.75 (0.697)	8.67 (0.896)	0.778
II;5;18	7.56 (0.781)	8.82 (0.911)	0.857
II;6;18	8.11 (0.838)	8.96 (0.926)	0.906
II;7;18	8.51 (0.879)	9.07 (0.937)	0.937
II;8;18	8.79 (0.908)	9.18 (0.948)	0.958
II;9;18	9.00 (0.930)	9.27 (0.958)	0.971
II;10;18	9.17 (0.947)	9.35 (0.966)	0.980
II;11;18	9.29 (0.960)	9.42 (0.973)	0.986
II;12;18	9.36 (0.967)	9.46 (0.977)	0.989
II;13;18	9.46 (0.977)	9.52 (0.983)	0.993
II;14;18	9.52 (0.983)	9.56 (0.988)	0.995
II;15;18	9.57 (0.989)	9.60 (0.992)	0.997
II;16;18	9.61 (0.993)	9.63 (0.995)	0.998
II;17;18	9.65 (0.997)	9.66 (0.998)	0.999

It was shown, for this set of item parameters, that there is no difference in the information between CML and MML estimation in complete designs. And the first interesting observation from Table 3.2 is that in any incomplete design, there is, irrespective of the estimation method, at least some loss of information compared to a complete testing design. For both estimation methods, CML and MML, the information increases with the test booklet length. It can be seen that the loss of information due to the design with CML estimation is quite large if we have a small number of items in the booklets. This is not true in MML estimation, where the loss due to the design with a few items is also rather small. The efficiency compared to a complete design is already at 0.855 with test length

of two items. However, the difference in the efficiency between CML and MML vanishes rather quickly with increasing test booklet length. The efficiency is already above 0.95 with an 8-item test booklet. From test lengths of 12 items or more there is hardly any difference in the loss due to the design with CML and MML estimation.

The comparisons for the item bank with items with different difficulty parameters are given in Table 3.3.

Table 3.3. Information comparison for item interlaced anchoring designs:  
 $\beta_i = -2.25 (0.25) 2.25, i = 1, \dots, 18$ .

design	CML-det (% of complete)	MML-det (% of complete)	INFEFF( $\beta; f(D); p_m(D), C_1$ )
complete	7.15	7.16	0.999
II;2;18	2.32 (0.324)	6.18 (0.863)	0.377
II;3;18	3.80 (0.531)	6.35 (0.887)	0.598
II;4;18	4.73 (0.662)	6.48 (0.905)	0.730
II;5;18	5.35 (0.748)	6.58 (0.919)	0.813
II;6;18	5.78 (0.808)	6.67 (0.932)	0.868
II;7;18	6.10 (0.853)	6.74 (0.941)	0.905
II;8;18	6.34 (0.887)	6.80 (0.950)	0.932
II;9;18	6.52 (0.912)	6.86 (0.958)	0.951
II;10;18	6.66 (0.931)	6.91 (0.965)	0.964
II;11;18	6.77 (0.947)	6.96 (0.972)	0.974
II;12;18	6.86 (0.959)	6.99 (0.976)	0.981
II;13;18	6.93 (0.969)	7.03 (0.982)	0.986
II;14;18	6.99 (0.978)	7.06 (0.986)	0.990
II;15;18	7.04 (0.985)	7.09 (0.990)	0.993
II;16;18	7.08 (0.990)	7.11 (0.993)	0.996
II;17;18	7.12 (0.996)	7.13 (0.996)	0.997

Note that, in this situation, there is a small loss of information if CML instead of MML estimation is used in a complete testing design. Compared to the results in Table 3.2, where all item parameters are equal, it can be seen that, for all

designs, the information efficiency of CML estimation compared to MML estimation is a little bit lower. It is true that in CML the efficiency compared to a complete design is always a bit larger if the item parameters are equal than if there are different item parameters. For MML, the differences between the two item sets are even smaller. However for MML, for booklets longer than 10 items the same trend is seen as with CML, while for shorter booklets the differences are in the opposite direction. For example, in the II;4;18 design, CML compared to a complete design gives 0.838 for equal parameters and for different parameters we have 0.808. For MML, we have in that same design 0.926 for equal parameters and 0.932 for unequal parameters. However, in the II;12;18 design MML shows 0.978 for all  $\beta = 0$  and 0.976 and for differing  $\beta$  for the efficiency compared to a complete design.

However, the general picture is the same: with very few items MML is more efficient, but this advantage disappears quickly with longer test booklets.

#### Results for block-interlaced anchoring designs

Table 3.4 and 3.5 give the results for 18 items with  $\beta_i = 0$  and for 18 items with  $\beta_i = -2.25 (0.25) 2.25, i=1, \dots, 18$  respectively. The efficiency of an incomplete design compared to a complete design is reported in parentheses in Tables 3.4 and 3.5. Two designs in these tables are the same as an item-interlaced design and a balanced block design respectively. This is indicated between brackets in the first column of the tables.

*Table 3.4. Information comparison for block-interlaced designs; 18 items with  $\beta_i = 0$ .*

design	CML-det (% of complete)	MML-det (% of complete)	INFEFF( $\beta; f(D); p_m(D), C_1$ )
complete	9.68	9.68	1
BI;2;18 (II)	3.40 (0.351)	8.28 (0.855)	0.411
BI;4;9	6.53 (0.675)	8.65 (0.894)	0.755
BI;6;6	7.95 (0.821)	8.93 (0.923)	0.891
BI;12;3 (BB)	9.36 (0.967)	9.46 (0.977)	0.989

Table 3.5. Information comparison for block-interlaced anchoring designs;  
 $\beta_i = -2.25 \text{ (0.25) } 2.25, i = 1, \dots, 18$ .

design	CML (% of complete)	MML-det (% of complete)	INFEFF( $\beta; f(D); p_m(D), C_1$ )
complete	7.15	7.16	0.999
BI;2;18 (II)	2.32 (0.324)	6.18 (0.863)	0.377
BI;4;9	4.61 (0.645)	6.47 (0.904)	0.713
BI;6;6	5.71 (0.799)	6.66 (0.930)	0.858
BI;12;3 (BB)	6.86 (0.959)	6.99 (0.976)	0.981

We see the same trends in the results for the block-interlaced designs as in the item-interlaced designs. It is seen in Table 3.4 and 3.5 that in all cases there is loss of information due to any incomplete design. Again, this loss is substantial for CML but not for MML, for very short test lengths. But for longer tests (larger than 6), the differences between CML and MML are relatively small. It can again be seen that in CML the efficiency compared to a complete design is always a bit larger in the case of equal item parameters than in the case of different item parameters. For MML, this is again only true for the designs with the longer booklets.

We see no large differences if we compare the results of item-interlaced and block-interlaced designs. However, if we compare the results for designs with equal booklet length in Table 3.2 and 3.4 (and in Table 3.3 and 3.5), the comparison always goes in the same direction. For the design with the short booklets, the information in CML estimation and the information in MML estimation as well as in the CML versus MML efficiency is always a bit larger in item interlaced than in block interlaced designs. With 12 items in a booklet no differences are left.

#### Results for balanced block designs

The results of the information comparison in balanced block designs are given



in Table 3.6 and 3.7. Two designs in these tables are the same as an item-interlaced design and a block-interlaced design respectively. This is indicated between brackets in the first columns of the tables.

Table 3.6. Information comparison for balanced block designs;  $\beta_i = 0, i = 1, \dots, 18$ .

design	CML-det (% of complete)	MML-det (% of complete)	INFEFF( $\beta; f(D); p_m(D), C_1$ )
complete	9.68	9.68	1
BB;(17.1);18 <sub>(II)</sub>	9.65 (0.997)	9.66 (0.998)	0.999
BB;(8.2);9	9.61 (0.993)	9.63 (0.995)	0.998
BB;(5.3);6	9.57 (0.989)	9.60 (0.992)	0.997
BB;(2.6);3 <sub>(BI)</sub>	9.36 (0.967)	9.46 (0.977)	0.989
BB;(5.2);18	9.22 (0.952)	9.37 (0.968)	0.983
BB;(3.3);10	9.07 (0.937)	9.29 (0.960)	0.976
BB;(4.2);18	8.94 (0.924)	9.23 (0.953)	0.969
BB;(3.2);15	8.33 (0.860)	8.99 (0.929)	0.927

All the results in Table 3.6 and 3.7 confirm the earlier findings. Although all the differences are very small, the differences and trends reported with the II and the IB designs are found again. The general result is that both estimation methods, CML and MML, are quite efficient in balanced block designs. Because of the ordering in the tables, it is easily checked that with increasing test length the information in CML, in MML, and also in the efficiency of CML compared to MML is increasing.

Table 3.7 Information comparison balanced block designs;

 $\beta_i = -2.25 (0.25) 2.25, i = 1, \dots, 18$ 

design	CML-det (% of complete)	MML-det (% of complete)	INFEFF( $\beta; f(D); p_m(D), C_1$ )
complete	7.15	7.16	0.999
BB;(17.1);18 <sub>(II)</sub>	7.12 (0.996)	7.13 (0.996)	0.997
BB;(8.2);9	7.08 (0.990)	7.11 (0.993)	0.996
BB;(5.3);6	7.04 (0.985)	7.09 (0.990)	0.993
BB;(2.6);3 <sub>(BI)</sub>	6.86 (0.959)	6.99 (0.976)	0.981
BB;(5.2);18	6.70 (0.937)	6.91 (0.965)	0.969
BB;(3.3);10	6.56 (0.917)	6.86 (0.958)	0.956
BB;(4.2);18	6.42 (0.898)	6.81 (0.951)	0.944
BB;(3.2);15	5.87 (0.821)	6.66 (0.930)	0.881

Finally, we can compare the results of the balanced block designs with the item-interlaced designs. The differences between the designs are very small. Only for designs with booklets with 10 items or less, it is seen that with booklets of the same length the CML and the MML information and the CML-MML information efficiency is larger for balanced block designs than for item-interlaced designs. And because the item-interlaced designs did a bit better than the block-interlaced designs, in general the balanced block designs can be preferred.

### 3.6.3 Results for a test taker as unit of cost

If the cost function  $C_2$ , defined in section 3.5, is used in the efficiency comparisons, we have a unit cost per test taker. The results can be used for comparing different designs with the same estimation method. Using this cost function, the results are easily translated to the number of persons needed in applying the design in order to get the same information as compared to a number of persons in a standard design. In Table 3.8, for the item bank with  $\beta_i = -2.25 (0.25) 2.25, i = 1, \dots, 18$ , for both CML and MML the results are given.

In both situations the standard of the information gathered in a complete design with a sample of 1000 students is taken.

Table 3.8. Efficiency of designs with cost function  $C_2$  and the item bank with  $\beta_i = -2.25 (0.25) 2.25, i = 1, \dots, 18$ .

Design	CML sample size	MML sample size
complete	1000	1000
II;2;18	27737	10427
II;3;18	11289	6765
II;4;18	6802	4972
II;5;18	4811	3917
II;6;18	3711	3220
II;7;18	4014	2731
II;8;18	2537	2369
II;9;18	2193	2087
II;10;18	1932	1865
II;11;18	1728	1683
II;12;18	1563	1536
II;13;18	1428	1410
II;14;18	1315	1303
II;15;18	1218	1211
II;16;18	1136	1132
II;17;18	1063	1063
BI;2;18 (II)	27737	10427
BI;4;9	6979	4979
BI;6;6	3756	3225
BI;12;3 (BB)	1563	1536
BB;(17.1);18(II)	1063	1063
BB;(8.2);9	1136	1132
BB;(5.3);6	1218	1211
BB;(2.6);3 (BI)	1563	1536
BB;(5.2);18	1920	1865
BB;(3.3);10	2179	2087
BB;(4.2);18	2505	2365
BB;(3.2);15	3654	3225

The general trends in the results in comparing the designs with the cost function  $C_2$  are the same as reported in section 3.6.2 for the cost function  $C_1$ . For both CML and MML, the complete design is most efficient and furthermore it is clear that with a constant cost per test taker it is better to have more items per test taker. Interesting in the results, is the direct translation of the efficiency in the number of observations which can be very useful in planning designs in practice. Although, the results in Table 3.8 are quite obvious, one example will be given for illustration. If we are planning to have booklets with 6 items, in CML we need 3711 persons in an item-interlaced design, 3756 in an block interlaced design and 3654 in a balanced block design to get the same information as in a complete design with 1000 persons. This indicates a clear preference for a balanced block design in the CML case. For MML, the needed numbers are 3220, 3225 and 3225 respectively, which indicates hardly any difference between the designs.

### **3.7 Conclusion**

In this study the comparison of the efficiency of CML and MML estimation of the item parameters of the Rasch model in incomplete designs was studied. The use of the concept of F-information (Eggen, 2000) was generalized to incomplete testing designs. It was shown that the only case in which CML and MML are equally efficient is when all item parameters are equal.

The standardized determinant of the F-information matrix is used for a scalar measure of information with respect to set of item parameters. In this chapter, the relation between the normalization of the Rasch model and this determinant was clarified. It was shown that the value of the determinants depends on a specific normalization. But comparing estimation methods with the defined information efficiency was shown to be independent of the chosen normalization.

Information comparisons were carried out with two item banks, two different cost functions of data collection and three different types of incomplete designs: item-interlaced, block-interlaced, and balanced block designs. The first general observation was that for both CML and MML some information is lost in all

incomplete designs compared to complete designs. Also the efficiency of CML compared to MML is in an incomplete design always smaller than in a complete design. The general trend is that with increasing test booklet length the information in CML, in MML, and also in the efficiency of CML compared to MML is increasing. The main differences between CML and MML were seen in relation to the length of the test booklet. It was demonstrated that with very small booklets (4 or less items), there is a substantial loss in information (more than 35%) with CML estimation, while this loss is only about 10% in MML estimation. However, with increasing test length, the differences between CML and MML quickly disappear. With test booklets with 12 items, the lowest reported efficiency of CML compared to MML was 0.979.

Between the results of the two considered sets of items no large differences were reported. The CML-MML efficiencies are somewhat higher with equal item parameters. Also the CML efficiency compared to a complete design is higher in case of equal parameters, while this is not found for all designs for the MML efficiency compared to a complete design.

No large differences were found between the design types studied. Their main results were the same. A slight advantage was seen for balanced block designs compared to item-interlaced designs, while these designs are expected to perform a bit better than block-interlaced designs. The differences between the designs were a bit larger for CML than for MML.

### 3.8 References

- Bhapkar, V.P. (1989). Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *Journal of Statistical Planning and Inference*, 21, 139-160.
- Cochran, W.G. & Cox, G.M. (1957). *Experimental designs. Second edition*. New York: Wiley.
- Eggen, T.J.H.M. (2000). On the loss of information in Conditional maximum likelihood estimation of item parameters. *Psychometrika*, 65, 337-362.
- Fischer, G.H. (1974). *Einführung in die Theorie der psychologischer Tests*. [Introduction to mental test theory.] Bern: Huber.
- Molenaar, I.W. (1995). Estimation of item parameters. In: Fischer, G.H. & Molenaar, I.W. (Eds.). *Rasch Models* (pp.39-51). New York: Springer.
- Pukelsheim, F. (1993). *Optimal design of experiments*. New York: Wiley.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Vale, C.D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, 10, 333-344.

## Appendix chapter 3

### *Proof Lemma 1*

Because the principal minors do not change if rows and columns of the matrix  $A$  are permuted in the same way, it is sufficient to prove that  $\det A^{(1)} = \det A^{(2)}$ .

Define two  $(k-1) \times (k-1)$ -matrices  $P$  and  $D$  by:

$$P = \begin{pmatrix} 1 & \mathbf{1}_{(k-2)}^T \\ \mathbf{0}_{(k-2)} & I_{k-2} \end{pmatrix} \text{ and } D = \text{diag}(-1, \mathbf{1}_{(k-2)}^T).$$

Then it is easily checked that  $\det P = 1$  and  $\det D = -1$ .

Furthermore, it yields that, because  $A$  is double centered (17)

$$A^{(1)} = DPA^{(2)}P^TD^T.$$

And the result follows directly.  $\square$

### *Proof Lemma 2*

Define  $B^* = K^T A^* = [b \ B]$

where  $b = \alpha c + a$

and  $B = ca^T + A$ .

Because  $a = -A\mathbf{1}$ , it follows that

$$B = -c\mathbf{1}^T A + A = A(I - c\mathbf{1}^T), \text{ from which it follows that}$$

$$\det B = \det A \cdot \det(I - c\mathbf{1}^T).$$

Because  $a = -A\mathbf{1}$  and  $\alpha = -\mathbf{1}^T a = \mathbf{1}^T A\mathbf{1}$ , it follows that

$$B\mathbf{1} = -b,$$

which means that the matrix  $B^*$  is row-centered, i.e the elements of each row sum to zero. Now

$$K^T A^* K = B^* K = (K^T B^{*T})^T \text{ where } B^{*T} = \begin{bmatrix} b^T \\ B^T \end{bmatrix} \text{ is column-centered.}$$

By that it follows that

$$\det(K^T A^* K) = \det B^T \cdot \det(I - c\mathbf{1}^T) = \det A \cdot (\det(I - c\mathbf{1}^T))^2. \square$$

## Chapter 4

### Item calibration in incomplete testing designs<sup>1</sup>

---

<sup>1</sup>This chapter is based on work which was conducted in cooperation with N.D. Verhelst. Parts of it were published as Measurement and Research Department Report 92-3, Cito:Arnhem.



## **Abstract**

This study discusses the justifiability of item parameter estimation in incomplete testing designs in item response theory. Marginal maximum likelihood (MML) as well as conditional maximum likelihood (CML) procedures are considered in three commonly used incomplete designs: random incomplete, multistage testing and targeted testing designs. Mislevy and Wu (1988) and Mislevy and Sheenan (1989) have shown that in these designs the justifiability of MML can be deduced from Rubin's (1976) general theory on inference in the presence of missing data. Their results are recapitulated. In this study it is shown that for CML estimation the justification must be established in an alternative way, by considering the neglected part of the complete likelihood. Incorrect uses of standard MML- and CML-algorithms are discussed.

## **4.1 Introduction**

Within the framework of item response theory (IRT) calibration and measurement designs often are incomplete designs. In test construction, item banking and equating studies the researcher frequently decides to administer only subsets of the total available item pool to the available (sampled) students. Sometimes there are just practical reasons for using incomplete designs, for example because of limited testing time not all the available items can be administered to every student. However, often efficiency is the motivating factor for building incomplete designs. Efficiency in item calibration is gained when (a priori) knowledge about the difficulty of the items and the ability of the students is used in allocating students to subsets of items ( e.g., Lord, 1980).

Algorithms for item calibration which allow for incomplete testing designs are implemented in several computer programs. For example, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), uses the marginal maximum likelihood (MML) approach in the one-, two- and three-parameter logistic test model and OPLM (Verhelst, Glas, & Verstralen, 1995), uses conditional maximum likelihood (CML) as well as MML procedures in general one-parameter logistic models. The application of these or similar computer programs in item banking, multistage testing, adaptive testing and equating studies is common psychometric practice. In these applications, however, some problems with incomplete designs are not generally recognized.

In this study calibration procedures in incomplete testing designs are reviewed. For convenience the one-parameter logistic test model for dichotomously scored items (Rasch, 1980) will be used for illustrative purposes. After reviewing IRT item parameter estimation in general, Rubin's (1976) concepts and theory on inference in the presence of missing data are summarized. Next, the applicability of this theory in MML as well as CML item calibration will be discussed. This will be elaborated for three commonly used incomplete design structures. For MML estimation, Mislevy and Wu (1988), with an emphasis on the estimation of person parameters and Mislevy and Sheenan (1989), focussing on the use of collateral information, have used the approach as presented here. The MML

results are essentially recapitulations of their work. In this chapter similar results for CML estimation of the item parameters are deduced via a different approach.

## 4.2 Item response theory

In IRT we consider the random vector, the response pattern  $\mathbf{X} = (X_{ij})$ ,  $i = 1, \dots, n$ ;  $j = 1, \dots, k$ , where  $X_{ij}$  is the response of student  $i$  to item  $j$ . With dichotomously scored items  $X_{ij} = 1$  if the answer is correct and  $X_{ij} = 0$  if the answer is not correct.

The one-parameter logistic model has as its basic equation (Rasch, 1980)

$$P(X_{ij} = x_{ij}) = \frac{\exp[(\theta_i - \beta_j)x_{ij}]}{1 + \exp(\theta_i - \beta_j)} = P_{\theta_i, \beta_j}(x_{ij}), \quad (1)$$

where  $x_{ij} \in \{0, 1\}$ ,  $i \in \{1, \dots, n\}$  and  $j \in \{1, \dots, k\}$ .

The distribution of  $X_{ij}$ , denoted by  $P_{\theta_i, \beta_j}(x_{ij})$ , follows the binomial distribution in which  $\theta_i$  is the ability parameter of student  $i$  and  $\beta_j$  the difficulty parameter of item  $j$ .

IRT models further assume local independence between item responses: given  $\theta_i$ , the item responses are independent. Furthermore, it is assumed that given  $\theta_i$ , the response of a student is not dependent on any other characteristic of the student. If we denote  $\mathbf{X}_i = (X_{ij}), j = 1, \dots, k$  as the response vector of student  $i$  and  $\mathbf{Y}_i$  as any other measured characteristic of student  $i$  (possibly multivariate but not functionally dependent on  $\mathbf{X}_i$ ):

$$P_{\theta_i, \beta}(\mathbf{x}_i | \mathbf{y}_i) = P_{\theta_i, \beta}(\mathbf{x}_i) = \prod_j P_{\theta_i, \beta_j}(x_{ij}), \quad (i = 1, \dots, n), \quad (2)$$

in which  $\beta = (\beta_j), j = 1, \dots, k$ . Finally independence of item responses between students is assumed, which means (with  $\theta = (\theta_i), i = 1, \dots, n$ ),

$$P_{\theta, \beta}(\mathbf{x}) = \prod_i P_{\theta_i, \beta}(\mathbf{x}_i) = \prod_i \prod_j P_{\theta_i, \beta_j}(x_{ij}). \quad (3)$$

Calibrating an item pool involves estimating the item parameters  $\beta$  and testing the validity of the model. In IRT maximum likelihood estimation is common, that is the probability of the particularly observed response pattern  $\mathbf{X} = \mathbf{x}$ , or the likelihood function

$$L(\beta, \theta; \mathbf{x}) = P_{\theta, \beta}(\mathbf{x}) \quad (4)$$

is maximized with respect to the parameters  $\beta$  and  $\theta$ . It is well known that because of the incidental parameters  $\theta_i$  in the model this does not lead to consistent estimates of the parameters, but in general two approaches are known to avoid this problem: CML and MML estimation.

#### 4.2.1 Conditional maximum likelihood estimation

If it is possible to construct a sufficient statistic  $S(\mathbf{X}_i)$  for the incidental parameter  $\theta_i$  (in the presence of the item parameter  $\beta$ ) we can factor the probability of the response pattern as

$$P_{\theta, \beta}(\mathbf{x}) = \prod_i P_{\beta}(\mathbf{x}_i \mid s(\mathbf{x}_i)) \cdot P_{\theta, \beta}(s(\mathbf{x}_i)), \quad (5)$$

in which  $S(\mathbf{X}) = (S(\mathbf{X}_i)), i = 1, \dots, n$  is the random vector of sufficient statistics for the ability parameters  $\theta$ . In (5), the first factor  $\prod_i P_{\beta}(\mathbf{x}_i \mid s(\mathbf{x}_i))$  does not depend on the ability parameters. And in CML estimation we proceed estimating the item parameters by just maximizing the conditional likelihood function with respect to  $\beta$ , which is the simultaneous conditional probability of the observed responses  $\mathbf{x}$ :

$$L_c(\beta; \mathbf{x} \mid s(\mathbf{x})) = \prod_i P_{\beta}(\mathbf{x}_i \mid s(\mathbf{x}_i)). \quad (6)$$

In CML estimation of the item parameters only random variations of the observations, fixing (given) the values of the conditioning statistics  $s(\mathbf{x}_i)$  are considered. The justification of this depends on whether all random variation that is relevant to the problem (here estimating the item parameters  $\beta$ ) is in this reduced frame of reference. This is easily seen to be heavily dependent on the

properties of the neglected part of (5). If the distribution of the sufficient statistic  $S(\mathbf{X}_i)$  would be completely independent of the item parameters  $\boldsymbol{\beta}$ , the justification would be obvious. However this condition is not fulfilled in our situation. But discarding this term is justified because Andersen (1973) has shown that the resulting CML estimators of  $\boldsymbol{\beta}$  are, under mild regularity conditions, consistent, and asymptotically normally distributed and efficient. Furthermore, in Eggen (2000) it was shown that the possible loss of information in CML estimation, by neglecting the information on  $\boldsymbol{\beta}$  in the distribution of  $S(\mathbf{X})$ , is very small already at short test lengths. (See also chapter 2 and 3 of this dissertation.) A major feature of CML estimation of the item parameters is that it is valid (i.e., having the above statistical properties) irrespective of any assumptions on the distribution of the ability of the students taking the test. The individual parameters are only part of the factor in the total likelihood which is justified to be neglected.

#### 4.2.2 Marginal maximum likelihood estimation

In MML estimation, model (3) is extended by assuming that the ability parameters  $\theta_i$  are a random sample from a population with probability density function given by  $g_\gamma(\theta)$ , with  $\gamma$  the (possibly vector valued) parameter of the ability distribution. Thus the response pattern  $\mathbf{X}$  as well as the ability  $\theta$  are considered random variables here. The  $\theta_i$  are not as before individual person ability parameters, but realizations of the unobservable random variable  $\theta$ . In MML we consider the marginal distribution of the response pattern  $\mathbf{X}$ ,

$$P_{\boldsymbol{\beta}, \gamma}(\mathbf{x}) = \int P_{\boldsymbol{\beta}, \gamma}(\mathbf{x}, \theta) d\theta = \prod_i \int P_{\boldsymbol{\beta}}(\mathbf{x}_i | \theta_i) g_\gamma(\theta_i) d\theta_i, \quad (7)$$

where  $P_{\boldsymbol{\beta}, \gamma}(\mathbf{x}, \theta)$  is the simultaneous distribution of the response pattern  $\mathbf{X}$  and the ability  $\theta$ .  $P_{\boldsymbol{\beta}}(\mathbf{x}_i | \theta_i) = \prod_j P_{\boldsymbol{\beta}_j}(x_{ij} | \theta_i)$  is the IRT model as in (3), giving the probability of a response vector of person  $i$ , with ability  $\theta_i$ .

In MML estimation the item parameters  $\boldsymbol{\beta}$  are simultaneously estimated with the parameter  $\gamma$  of the ability distribution by maximizing the marginal

probability of the observed response pattern  $\mathbf{x}$  (the marginal likelihood function) with respect to the parameters, that is,

$$L_m(\beta, \gamma; \mathbf{x}) = \prod_i \int P_{\beta}(\mathbf{x}_i | \theta_i) \cdot g_{\gamma}(\theta_i) d\theta_i. \quad (8)$$

The consistency of the item parameter estimators with MML can be deduced from the work by Kiefer and Wolfowitz (1956). In practice, the most popular approach here is to assume that the ability distribution of  $\theta$  is normal with  $\gamma = (\mu, \sigma^2)$ . Bock and Aitkin (1981) were the first to give computational procedures for maximizing (8) using the EM-algorithm.

### 4.3 Inference and missing data

Rubin (1976) and Little and Rubin (1987) present a general framework for inference in the presence of missing data. Here their defined concepts and some of the results are summarized. First, some notations and definitions.

Let  $\mathbf{U} = (U_1, \dots, U_m)$  a vector random variable with probability density function  $f_{\tau}(\mathbf{u})$ .  $\tau$  is a vector parameter, on which we want to draw inferences on the basis of the data, a sample realization  $\mathbf{u}$ . Assume for convenience  $m = n.k$ , with  $k$  the number of variables and  $n$  the number of persons sampled. In the presence of missing data a vector random design variable, or missing data indicator,  $\mathbf{M} = (M_1, \dots, M_m)$  is defined, indicating whether a variable  $U_j$ , is actually observed,  $m_j = 1$ , or not observed,  $m_j = 0$ . The observed value of  $\mathbf{M}$  ( $\mathbf{m}$ ) effects a partition of the vector random variable  $\mathbf{U}$  and of its observed value:  $\mathbf{U} = (\mathbf{U}_{obs}, \mathbf{U}_{mis})$  and  $\mathbf{u} = (\mathbf{u}_{obs}, \mathbf{u}_{mis})$ . The sets of indices of observed and not observed variables are  $obs = \{j | m_j = 1\}$  and  $mis = \{j | m_j = 0\}$ .

In Rubin's (1976) theory the conditional distribution of the missing data indicator given the data has a key role:

$$P_{\phi}(\mathbf{M} = \mathbf{m} | \mathbf{U} = \mathbf{u}) = h_{\phi}(\mathbf{m} | \mathbf{u}), \quad (9)$$

which is defined as the distribution corresponding to the process that causes the missing data, with  $\phi$  a possibly vector valued parameter. In general  $\phi$  can be dependent on the parameter of interest  $\tau$ : they could have common or functionally related elements.

The general problem in inference in the presence of missing data is that we have a sample realization of  $\mathbf{M}$  and  $\mathbf{U}_{obs}$  and we want to infer on the parameter  $\tau$  of the distribution of the only partially observed  $\mathbf{U}$ . In the presence of missing data, the basis for inference on  $\tau$  should be the joint distribution of  $\mathbf{M}$  and  $\mathbf{U}_{obs}$ :

$$\int_{\mathbf{u}_{mis}} f_{\tau, \phi}(\mathbf{u}, \mathbf{m}) d\mathbf{u}_{mis} = \int_{\mathbf{u}_{mis}} f_{\tau}(\mathbf{u}) \cdot h_{\phi}(\mathbf{m} | \mathbf{u}) d\mathbf{u}_{mis} . \quad (10)$$

Because we are only interested to infer on the parameter  $\tau$  of the distribution of the partially observed  $\mathbf{U}$ , a possible approach could be to ignore in the inference the process that causes the missing data. Following Rubin (1976), ignoring the process that causes missing data means:

- (a) fixing the random variable  $\mathbf{M}$  at the observed pattern of missing data  $\mathbf{m}$  and
- (b) assuming that the values of the observed data  $\mathbf{u}_{obs}$  are realizations of the marginal density of  $\mathbf{U}_{obs}$ :

$$\int_{\mathbf{u}_{mis}} f_{\tau}(\mathbf{u}) d\mathbf{u}_{mis} . \quad (11)$$

When we ignore the process that causes the missing data, not all possible random variation in the data due to sampling of  $\mathbf{M}$  and  $\mathbf{U}_{obs}$  is considered, but only random variation due to  $\mathbf{U}_{obs}$  fixing the random variable  $\mathbf{M}$  at the particularly observed pattern  $\mathbf{m}$ . The generally more convenient form (11) is used instead of (10) in the inference on  $\tau$ . This is illustrated in the following example.

### Example 1

Suppose we partially observe the responses of  $n$  persons on one item ( $j$ ) following the Rasch model (1), and we write  $p_{\beta_j}(\theta_i)$  for correctly answering item  $j$  by person  $i$ . If we deduce the distribution of the observed responses from the

joint distribution of the response variable and the missing data indicator (10), then we have:

$$\begin{aligned} \prod_{i=1}^n p_{\beta_j}(\theta_i)^{x_{ij}} (1-p_{\beta_j}(\theta_i))^{(1-x_{ij})} h_{\phi}(m_{ij}|x_{ij}) = \\ \prod_{i \in obs} [p_{\beta_j}(\theta_i)^{x_{ij}} (1-p_{\beta_j}(\theta_i))^{(1-x_{ij})} h_{\phi}(m_{ij}=1|x_{ij})] \cdot \\ \prod_{i \in mis} [p_{\beta_j}(\theta_i) h_{\phi}(m_{ij}=0|x_{ij}=1) + (1-p_{\beta_j}(\theta_i)) h_{\phi}(m_{ij}=0|x_{ij}=0)]. \end{aligned} \quad (12)$$

If we ignore the process that causes the missing data, the marginal distribution of the observed responses from (11) would be:

$$\begin{aligned} \prod_{i=1}^n p_{\beta_j}(\theta_i)^{x_{ij}} (1-p_{\beta_j}(\theta_i))^{(1-x_{ij})} = \\ \prod_{i \in obs} p_{\beta_j}(\theta_i)^{x_{ij}} (1-p_{\beta_j}(\theta_i))^{(1-x_{ij})} \cdot \prod_{i \in mis} \sum_{x_{mis}} p_{\beta_j}(\theta_i)^{x_{ij}} (1-p_{\beta_j}(\theta_i))^{(1-x_{ij})} \end{aligned} \quad (13)$$

In the right hand side of (13), in the second factor the summation is taken over all possible values of  $x_{mis}$  (0 or 1), and it will be clear that this factor equals 1 and the marginal distribution of the responses is simply:

$$\prod_{i \in obs} p_{\beta_j}(\theta_i)^{x_{ij}} (1-p_{\beta_j}(\theta_i))^{(1-x_{ij})}. \square \quad (14)$$

It will be clear that ignoring the missing data process does not necessarily lead to a correct inference on  $\tau$ . Firstly, we possibly disregard the possible influence of  $\phi$  on  $\tau$ . Possible restrictions, due to  $\phi$ , are not taken in account in the inference on  $\tau$ . Secondly, it is understood that the data  $\mathbf{u}_{obs}$  are in fact no realizations of (11) but of the conditional density of  $\mathbf{U}_{obs}$  given the random variable  $\mathbf{M}$  took the fixed value  $\mathbf{m}$ :

$$\int_{\mathbf{u}_{mis}} f_{\tau, \phi}(\mathbf{u} | \mathbf{m}) d\mathbf{u}_{mis} = \int \frac{f_{\tau, \phi}(\mathbf{u}, \mathbf{m})}{f_{\phi}(\mathbf{m})} d\mathbf{u}_{mis} = \int \frac{f_{\tau}(\mathbf{u}) \cdot h_{\phi}(\mathbf{m} | \mathbf{u})}{\int f_{\tau}(\mathbf{u}) \cdot h_{\phi}(\mathbf{u} | \mathbf{m}) d\mathbf{u}} d\mathbf{u}_{mis}, \quad (15)$$

which is in general not equal to (11).



In Rubin (1976) sufficient conditions as well as necessary and sufficient conditions are specified such that ignoring the process that causes the missing data yields the correct inference about  $\tau$ . We will only consider the sufficient conditions in direct likelihood inference, because these suffice for our arguments. By direct likelihood inference is meant inference on parameter(s) based on comparison of likelihoods as e.g. the determination of a maximum likelihood estimator and likelihood ratio tests. The sufficient conditions are on the distribution  $h_{\phi}(\mathbf{m} \mid \mathbf{u})$ . Define:

1. The missing data are missing at random (MAR) if for each value of  $\phi$

$$h_{\phi}(\mathbf{m} \mid \mathbf{u}_{obs}, \mathbf{u}_{mis}) = h_{\phi}(\mathbf{m} \mid \mathbf{u}_{obs}) \text{ for all } \mathbf{u}_{mis}, \quad (16)$$

that is, the missingness of the data does not depend on the not observed values of  $\mathbf{U}_{mis}$ , but may depend on the observed values of  $\mathbf{U}_{obs}$ .

2. The missing data are missing completely at random (MCAR) if for each value of  $\phi$

$$h_{\phi}(\mathbf{m} \mid \mathbf{u}_{obs}, \mathbf{u}_{mis}) = h_{\phi}(\mathbf{m}) \text{ for all } \mathbf{u}_{mis} \text{ and } \mathbf{u}_{obs}. \quad (17)$$

Note that MCAR implies MAR.

3. The parameter  $\phi$  is distinct (D) from  $\tau$  if the joint parameter space of  $(\phi, \tau)$  is the cartesian product of the parameter space of  $\phi$  and the space of  $\tau$ .  
Distinctness means that all possible values of  $\phi$  are possible in combination with all possible values of  $\tau$ .

These three definitions enable us to state Rubin's (1976) ignorability principle: if both MAR and D hold ignoring the process that causes the missing data gives correct direct likelihood inferences about  $\tau$ .

This means that instead of using the full-likelihood

$$L(\tau, \phi; \mathbf{u}_{obs}, \mathbf{m}) = f_{\tau, \phi}(\mathbf{u}_{obs}, \mathbf{m}) = \int f_{\tau, \phi}(\mathbf{u}, \mathbf{m}) d\mathbf{u}_{mis}, \quad (18)$$

the simple likelihood function

$$L(\tau; \mathbf{u}_{obs}) = f_{\tau}(\mathbf{u}_{obs}) = \int f_{\tau}(\mathbf{u}) d\mathbf{u}_{mis}, \quad (19)$$

can be used for inferring on  $\tau$ . Ignoring the process that causes missing data is of course also justified if the stronger condition MCAR, instead of MAR, (and D) is met.

It is noted that these conditions only guarantee correct direct likelihood inferences as determining the correct maximum likelihood estimate. It is not guaranteed that the resulting estimates in using (18) or (19) will have the same statistical properties, such as consistency or asymptotic normality. In general, then stronger conditions have to be fulfilled (Rubin, 1976).

#### 4.4 Incomplete calibration designs

Using incomplete testing designs is very common in the application of IRT. Although many variants are possible, one of three calibration design structures is commonly used: random incomplete designs, multistage testing designs and targeted testing designs. The following notation and assumptions are used in describing these designs.

We have  $T$  test forms, indexed by  $t = 1, \dots, T$ . From the total item pool of  $k$  items, subsets of  $k_t, (t = 1, \dots, T)$  items are assembled in the test forms. We assume that there is overlap in items between the test forms. Via the linking items the item pool can be calibrated on the same scale. Fischer (1981) gives the exact conditions that have to be fulfilled for the existence and uniqueness of the item parameter estimates in incomplete designs using CML in the Rasch model. In practice, these conditions are almost always met if there are some common items in the test forms. In MML estimation the linking in incomplete designs is also mostly established via common items. Although, for MML estimation Glas (1989) has shown that in the special case where we do not have a linked design

but assume a common ability distribution for all sampled students the parameters are estimable.

We assume that every student takes only one test form and for every student taking items from the pool we define a design or item indicator vector with as many elements as there are items in the item pool ( $k$ ). The item indicator variable for every student  $\mathbf{R}_i$  can take  $T$  values:

$$\mathbf{r}_t = \text{perm}_t(\mathbf{1}_{k_t}, \mathbf{0}_{k-k_t}), \quad (t = 1, \dots, T). \quad (20)$$

Each value is a permutation of the vector  $(\mathbf{1}_{k_t}, \mathbf{0}_{k-k_t})$ , indicating that there are  $k_t$  values 1 at the elements indexed by the items in the administered test  $t$ , and  $k - k_t$  values 0.

It is noted that the missing data indicator  $\mathbf{M}$  is strongly related to the item indicator  $\mathbf{R}$ . In our applications it is always true that  $\mathbf{R} \subset \mathbf{M}$ . But  $\mathbf{R}$  concerns only the indication whether items are observed, while  $\mathbf{M}$  also concerns the observation or missingness of other variables considered in a problem. More specifically, when the ability  $\theta$  is considered as a random variable as in MML estimation (8), we will use the indicator variable  $\mathbf{M}$ , having a value zero for all realizations of  $\theta$ .

#### 4.4.1 Random incomplete designs

In random incomplete designs the researcher decides which test form is taken by which students without using any a priori knowledge on the ability of a student. Every student has an a priori known chance of taking one of the  $T$  test forms. In these designs the test forms are often assembled from the item pool in such a way that the forms have an equal number of items and are parallel in content and difficulty. A test form can be randomly assigned to a student so that every student has an equal chance of getting a particular test form. Or more generally a student gets a test form with a known probability  $\phi_t$  such that  $\sum_{t=1}^T \phi_t = 1$ . The distribution of the item indicator variable  $\mathbf{R}_i$  is given by:

$$P(\mathbf{R}_i = \mathbf{r}_t) = \phi_t \text{ with } t \in \{1, \dots, T\}, (i = 1, \dots, n). \quad (21)$$

#### 4.4.2 Multistage testing designs

In multistage testing designs the assignment of students to subsets of items from the total item pool in a testing stage is based on the observed responses in the former stage. A typical example is given in Figure 4.1. All students in the sample take the first stage test which is of medium difficulty. This (part of the) test is called the routing test. Students with high scores on the routing test are administered a more difficult subset of items from the pool in the next stage and students with low scores a more easy subset. The same procedure is possibly continued in next testing stages.

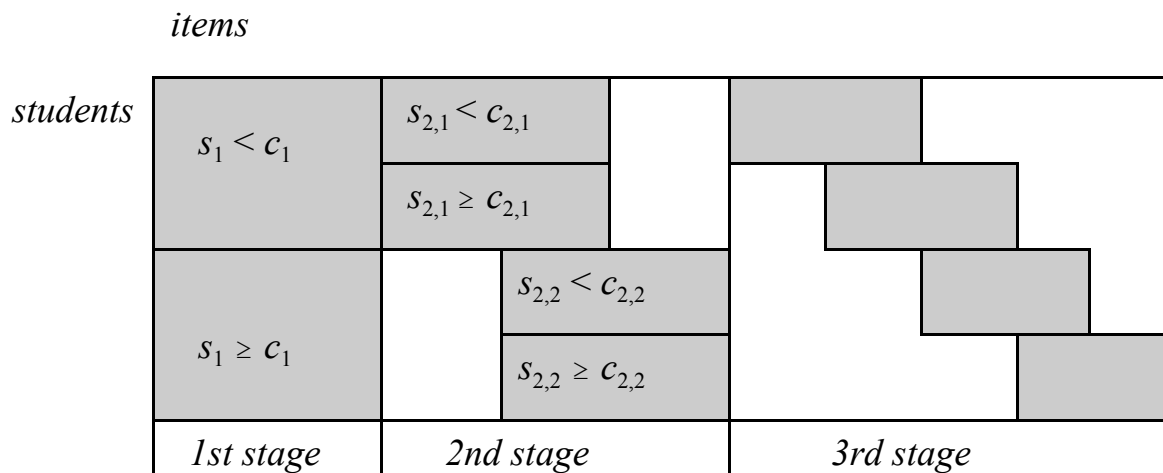


Figure 4.1: Example of a multistage testing design

In Figure 4.1,  $s_1$  indicates the score on the items of the first stage (routing) test, and  $s_{2,1}$  is the score on a second stage (routing) test which content depends on the score on the first routing test. In each stage the score is compared to a cut-off  $c.$ , on which it is decided which items are administered next. In this example, considering the total data matrix, the total number of tests  $T$  is 4.

Multistage testing was introduced (Lord, 1971) for efficiently measuring the ability of students, but it is understood that the underlying principle can also be applied in designs for the calibration of the items. A limiting case of multistage testing is computerized adaptive testing (see chapter 1 of this dissertation), where the stages have a length of only one item: after every item, the next item

administered is selected on the basis of the result on the previously administered items.

In a multistage testing design, as in a random incomplete design, the item indicator variable for every student  $\mathbf{R}_i$  can take as many values as there are tests  $T$  (see (20)). The distribution of  $\mathbf{R}_i$  has always the following form:

$$P(\mathbf{R}_i = \mathbf{r}_t \mid \mathbf{x}_{obs,i}) = \phi_t(\mathbf{x}_{obs,i}), \quad (t = 1, \dots, T), (i = 1, \dots, n). \quad (22)$$

If a function of observed item scores  $\mathbf{x}_{obs,i}$  meets a criterion for getting test  $t$ , the item indicator variable  $\mathbf{R}_i$  takes the value  $\mathbf{r}_t$  with probability  $\phi_t$ . If the criterion is not met the probability is  $1 - \phi_t$ . It should be understood that in a multistage testing design the probability of a certain design is not constant for all values of  $\mathbf{x}_{obs,i}$ , because in that case the design is random incomplete.

#### Example 2.

We have a routing test consisting of 3 items with  $\beta_1 = \beta_2 = \beta_3 = 0$ . With a total score of 0 or 1 on these 3 items an easier test of 4 items with parameters  $\beta_4 = -1.25$ ,  $\beta_5 = -1.0$ ,  $\beta_6 = -0.5$  and  $\beta_7 = 0.5$  is conducted. When the score on the routing test is 2 or 3, a harder test, having two items in common with the easier, with the parameters  $\beta_6 = -0.5$ ,  $\beta_7 = 0.5$ ,  $\beta_8 = 1.0$  and  $\beta_9 = 1.25$  is administered. The functions  $\phi_t(\mathbf{x}_{obs,i})$  (22) can then be defined as:  $\phi_1(\mathbf{x}_{obs,i}) = 1$  if  $\sum_{i=1}^3 x_{ij} \leq 1$ , and  $\phi_2(\mathbf{x}_{obs,i}) = 1$  if  $\sum_{i=1}^3 x_{ij} > 1$ , where test 1 is the easier test and test 2 the harder test.  $\square$

#### 4.4.3 Targeted testing designs

In targeted testing designs the structure of the design is determined a priori on the basis of background information, say values of a random variable  $\mathbf{Y}$  of the students. This background variable is usually positively related to the ability. Students with values of  $\mathbf{Y}$  which are expected to have lower abilities are administered easier test forms, and students with values of  $\mathbf{Y}$  expected to have higher abilities are administered the more difficult forms. As in multistage testing

designs gains in precision of the estimates are to be expected. An example of a variable often used in these designs is the grade level of the student.

We will assume that the variable  $\mathbf{Y}$  of the students is categorical (or categorized), taking (or distinguishing)  $T$  values:  $\mathbf{y}_1, \dots, \mathbf{y}_T$ . In targeted testing, for each value of  $\mathbf{Y}$  a different subset from the total item pool is administered to the students. The variable  $\mathbf{Y}$  can, besides for the assignment of the items to the students, also play a role in the sampling of the students. We can distinguish two situations. First, the background variable  $\mathbf{Y}$  is only used in the assignment of items or tests to students and not in the sampling of students. Second, the  $\mathbf{Y}$  is used in the sampling of students as well as in the assignment of tests to students.

In the first situation the role of using  $\mathbf{Y}$  is limited to increase the precision of the parameter estimates of the items to be calibrated. In this situation there is no explicit interest in the variable  $\mathbf{Y}$  itself. There is, for instance, no interest to have estimates of the parameters of the ability distribution for each distinguished level of  $\mathbf{Y}$ . Here the students are sampled from one population with no regard to the values of  $\mathbf{Y}$ . An example of an application of this form of targeted testing was used in the Dutch National Assessment program (Verhelst & Eggen, 1989), where the assignment of one of two possible tests was based on the judgement of the teacher.

In second situation the background variable also plays a role in sampling the students. In this case there is an explicit interest in the variable  $\mathbf{Y}$  itself. A situation often occurring is that  $\mathbf{Y}$  is the stratification variable in the sampling of students from the total population. Often the sampling proportions within the strata are not the same in the total population and one is explicitly interested in estimates of the ability distribution of the different strata and possibly, but not necessarily, in the total population. In this case, unlike in the first situation, the sampled students can in general not be considered to be a sample from one population but are from a total population divided in subpopulations of interest.

Where relevant we will distinguish these two targeted testing situations as (a) targeted testing with student sample from one population (TTOP), and as (b) targeted testing with student samples from multiple (sub)populations (TTMP).

In targeted designs the item indicator variable  $\mathbf{R}_i$  for every student can again take as many values as there are tests (see 18). The distribution of  $\mathbf{R}_i$  is given by

$$P(\mathbf{R}_i = \mathbf{r}_t \mid \mathbf{Y}_i = \mathbf{y}_t) = \phi_t(\mathbf{y}_t), \quad (t=1, \dots, T). \quad (23)$$

For any (distinguished) value of the background variable  $\mathbf{Y}$ , there is a fixed probability that a certain test is administered. An example is the gender of the student. A boy ( $\mathbf{y}_i = 1$ ) then gets with a probability  $\phi_1(\mathbf{y}_i = 1)$  test 1 and a girl with probability  $\phi_1(\mathbf{y}_i = 2)$ . Similar probabilities can be specified for a second test. In practice, often  $\phi_t = 1$ , which means that given the value  $\mathbf{y}_i$ , a specific test is administered. The formal resemblance between a targeted testing (23) and multistage testing design (22) is noted. But the difference is also clear: in a targeted testing design,  $\mathbf{Y}_i$  can be any measured characteristic of a student, with the exception that it is not (based on) responses to items whose parameters are to be estimated as we have in multistage testing (22).

#### 4.5 Item calibration and missing data

Although item calibration in incomplete testing designs is common in psychometric practice and modern computer programs can analyze incomplete designs, it is commonly assumed that the stochastic nature of the item indicator variable  $\mathbf{R}$  does not play a role in the calibration. In implemented computer algorithms the design variable value is fixed at the observed value and only random variations in the observed item responses are considered. One could say that the ignorability principle is assumed to hold. In this section we will explore the justifiability of this practice in the incomplete calibration designs described in the paragraph 4.4. We will treat marginal as well as conditional estimation of the item parameters in these designs. We assume that we have tested a group of  $n$  students, for which the observed and missing variables are notated with  $\mathbf{U}_{obs,i}$  and  $\mathbf{U}_{mis,i}$ ,  $i = 1, \dots, n$ ,  $\mathbf{U} = (\mathbf{U}_{obs}, \mathbf{U}_{mis})$ , with  $\mathbf{U}_{obs} = (\mathbf{U}_{obs,1}, \dots, \mathbf{U}_{obs,n})$  and  $\mathbf{U}_{mis} = (\mathbf{U}_{mis,1}, \dots, \mathbf{U}_{mis,n})$ . The missing data indicator is  $\mathbf{M} = (\mathbf{M}_1, \dots, \mathbf{M}_n)$ , in which

every element  $\mathbf{M}_i$  is a vector of the same length as there are variables (observed and unobserved).

#### 4.5.1 The marginal model and missing data

Mislevy and Wu (1988) and Mislevy and Sheenan (1989) have given a thorough treatment of the justifiability of using background information of students and of the MML estimation procedure in incomplete designs. They check the ignorability conditions for the design variable in incomplete designs and the conditions for background variable(s) and the design variable in targeted testing. Partly recapitulating their work, we will give next the results for complete, random incomplete, multistage and targeted testing designs.

##### MML in complete, random incomplete and multistage testing designs

First we note that the justification of using MML for complete data, see (7) and (8), can also be deduced from the general framework of Rubin for inference in the presence of missing data. Complete data MML can be described as a procedure in which we have missing data and the ignorability principle is applied in likelihood inference. This is readily seen as follows. The variable on which we want to base our inference on is  $\mathbf{U} = (\mathbf{X}, \boldsymbol{\theta}) = (\mathbf{X}_1, \boldsymbol{\theta}_1, \dots, \mathbf{X}_n, \boldsymbol{\theta}_n)$  in which  $\mathbf{X}_i$  is as before the random answer vector of student  $i$  on the  $k$  items administered. The parameter to be estimated is  $\boldsymbol{\tau} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ . In the complete data situation the  $\mathbf{X}_i$  are always observed and the  $\boldsymbol{\theta}_i$  are always missing. So we have for every student  $i$  a degenerated design distribution, that equals its item indicator distribution

$$P(\mathbf{M}_i = (\mathbf{1}_k, 0)) = P(\mathbf{R}_i = (\mathbf{1}_k)) = 1, \quad (i = 1, \dots, n). \quad (24)$$

The partition which the observed design variable  $\mathbf{m}_i$  effects is

$$\mathbf{U}_{obs,i} = \mathbf{X}_i \quad \text{and} \quad \mathbf{U}_{mis,i} = \boldsymbol{\theta}_i, \quad (i = 1, \dots, n). \quad (25)$$



Because the parameter space of the distribution of  $\mathbf{M}$  is empty and MCAR is clearly met, the marginal distribution of  $\mathbf{U}_{obs}$  (here  $\mathbf{X}$ ) can be used by the ignorability principle for correct likelihood inference:

$$\int_{\mathbf{u}_{mis}} f_{\tau}(\mathbf{u}) d\mathbf{u}_{mis} = \int P_{\beta, \gamma}(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \prod_i \int P_{\beta}(\mathbf{x}_i | \boldsymbol{\theta}_i) g_{\gamma}(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i. \quad (26)$$

Which is identical to (8).

In random incomplete designs and multistage testing designs the ignorability conditions are also fulfilled. In Table 4.1 we give for these designs and for the complete testing design respectively the observed and unobserved variables and the design distribution.

Table 4.1. Variables in incomplete testing designs

design	$\mathbf{U}_{obs,i}$	$\mathbf{U}_{mis,i}$	$h_{\phi}(\mathbf{m}_i   \mathbf{u}_{obs,i}, \mathbf{u}_{mis,i})$
complete	$\mathbf{X}_i$	$\boldsymbol{\theta}_i$	$P(\mathbf{M}_i = (\mathbf{1}_k, 0)) = P(\mathbf{R}_i = (\mathbf{1}_k)) = 1$
random incomplete	$\mathbf{X}_{obs,i}$	$\mathbf{X}_{mis,i}, \boldsymbol{\theta}_i$	$P(\mathbf{M}_i = (\mathbf{r}_t, 0)) = P(\mathbf{R}_i = \mathbf{r}_t) = \phi_t$
multistage	$\mathbf{X}_{obs,i}$	$\mathbf{X}_{mis,i}, \boldsymbol{\theta}_i$	$P(\mathbf{M}_i = (\mathbf{r}_t, 0)   \mathbf{x}_{obs,i}) = P(\mathbf{R}_i = \mathbf{r}_t   \mathbf{x}_{obs,i}) = \phi_t(\mathbf{x}_{obs,i})$

The design distribution in random incomplete and in multistage testing design follow respectively from (21) and (22). In random incomplete designs the MCAR condition is fulfilled and in multistage testing design the MAR condition. In both designs the D condition is clearly met. Therefore ignorability holds in these designs and MML can be applied using the marginal distribution of the observations. This can readily be checked by considering, e.g in the multistage testing design, the distribution of  $(\mathbf{U}_{obs}, \mathbf{M})$  needed for the full likelihood:

$$\begin{aligned}
\int_{\mathbf{u}_{mis}} P_{\tau,\phi}(\mathbf{u}, \mathbf{m}) d\mathbf{u}_{mis} &= \int_{\theta} \int_{\mathbf{x}_{mis}} P_{\beta,\gamma,\phi}(\mathbf{x}_{obs}, \mathbf{x}_{mis}, \theta, \mathbf{m}) d\mathbf{x}_{mis} d\theta = \\
\int_{\theta} \int_{\mathbf{x}_{mis}} P_{\beta,\gamma}(\mathbf{x}_{obs}, \mathbf{x}_{mis}, \theta) \cdot h_{\phi}(\mathbf{m} | \mathbf{x}_{obs}, \mathbf{x}_{mis}, \theta) d\mathbf{x}_{mis} d\theta &= \\
h_{\phi}(\mathbf{m} | \mathbf{x}_{obs}) \int_{\theta} P_{\beta,\gamma}(\mathbf{x}_{obs}, \theta) d\theta &= \\
\prod_i h_{\phi}(\mathbf{m}_i | \mathbf{X}_{obs,i}) \prod_i \int_{\theta_i} P_{\beta}(\mathbf{X}_{obs,i} | \theta_i) \cdot g_{\gamma}(\theta_i) d\theta_i .
\end{aligned} \tag{27}$$

In (27) the third equality holds because of MAR resulting in a factorization of the full likelihood in a term independent of  $(\beta, \gamma)$  and the marginal distribution of  $\mathbf{X}_{obs}$ . So just considering the marginal distribution of  $\mathbf{X}_{obs}$  will thus give the correct maximum likelihood estimates of  $\beta$  and  $\gamma$ .

Note that if we indicate by  $n_t$  the number of students taking test  $t$  with  $\sum_{t=1}^T n_t = n$  and define  $\beta_{(t)}$  as the  $k_t$ - vector of the item parameters of the items in test  $t$ , we can rewrite the second factor of (27) as

$$\int_{\theta_i} P_{\beta}(\mathbf{x}_{obs,i} | \theta_i) \cdot g_{\gamma}(\theta_i) d\theta_i = \prod_{t=1}^T \prod_{i=1}^{n_t} \int_{\theta_{(t)}} P_{\beta_{(t)}}(\mathbf{x}_{obs,i} | \theta_{(t)}) \cdot g_{\gamma}(\theta_{(t)}) d\theta_{(t)} . \tag{28}$$

The marginal likelihood in the incomplete design case is thus written as a product of  $T$  complete data marginal likelihoods.

#### MML in targeted testing designs

Mislevy and Sheenan (1989) present a more general discussion on the effect of using or not using (ignoring) the background information of the students in MML item calibration, depending on the role this variable  $\mathbf{Y}$  has in the testing design. In complete testing designs  $\mathbf{Y}$  can play a role in the sampling: students can be sampled from one population, or (stratified) from multiple subpopulations. In targeted testing, besides the possible sampling role of  $\mathbf{Y}$ , the background variable is used in the assignment of items to students. Mislevy and Sheenan

(1989) show that in all these situations the justifiability of MML item calibration can be deduced from Rubins missing data framework. We will reconsider their results.

Assume  $\mathbf{Y}$  to be a categorical (or categorized) variable taking one of  $L$  values, establishing a division of the total student population in  $L$  subpopulations. The value of  $\mathbf{Y}$  for student  $i$  is defined as  $\mathbf{y}_i = (y_{i1}, \dots, y_{iL})$  with  $y_{i\ell} = 1$  if student  $i$  is associated with subpopulation  $\ell$  and 0 if not,  $\ell = 1, \dots, L$ . If  $y_{i\ell} = 1$  we will alternatively write  $\mathbf{y}_i = \mathbf{y}^{(\ell)}$ . The ability distribution  $g_{\gamma}(\theta)$  of the total population in that case can be expressed as a finite mixture of  $L$  subpopulation ability distributions:

$$\begin{aligned} g_{\gamma}(\theta) &= \sum_{\ell=1}^L P(\theta, \mathbf{Y} = \mathbf{y}^{(\ell)}) = \sum_{\ell=1}^L P(\theta \mid \mathbf{Y} = \mathbf{y}^{(\ell)}) \cdot P(\mathbf{Y} = \mathbf{y}^{(\ell)}) \\ &= \sum_{\ell=1}^L g_{\gamma_{\ell}}(\theta) \cdot \pi_{\ell}, \end{aligned} \quad (29)$$

in which  $\gamma_{\ell}$  is the possibly vector valued parameter of the ability distribution in subpopulation  $\ell$  and  $\pi_{\ell}$  the proportion of subpopulation  $\ell$  in the total population.

In complete testing designs using or not using  $\mathbf{Y}$  in MML item calibration is equivalent with considering  $\mathbf{Y}$  as observed or missing data. In Eggen and Verhelst (1992) detailed checks of Rubins ignorability conditions are given, leading to the same results as Mislevy and Sheenan (1989). Summarized the results are: using  $\mathbf{Y}$  in MML item calibration makes it possible, independent of the sampling role, to estimate the item parameters  $\beta = (\beta_1, \dots, \beta_k)$  and the population parameters  $\gamma = (\gamma_1, \dots, \gamma_{\ell}, \dots, \gamma_L, \pi_1, \dots, \pi_{\ell}, \dots, \pi_L)$  simultaneously. The justifiability of ignoring  $\mathbf{Y}$  depends on the sampling role of  $\mathbf{Y}$  in the design: correct estimates of the item parameters and the population parameters in MML item calibration are guaranteed only, when we have a random sample from one population. In case we have samples from multiple subpopulations ignoring  $\mathbf{Y}$  possibly leads to wrong estimates. See Mislevy and Sheenan (1989) and Eggen and Verhelst (1992) for details.

In targeted testing designs two possible roles of the background variable  $\mathbf{Y}$  can be distinguished. First in TTOP,  $\mathbf{Y}$  has no role in the sampling of the students, but depending on the values of  $\mathbf{Y}$  different subsets of the item pool are administered. Second in TTMP  $\mathbf{Y}$  has both a role in the sampling of the students and in the assignment of items to students.

In TTOP we have a random sample from the total population with ability distribution  $g_{\mathbf{Y}}(\boldsymbol{\theta})$  (29). For students with value  $y^{(\ell)}$  of  $\mathbf{Y}_i$  denote with  $\boldsymbol{\beta}_{(\ell)}$  the  $k_{\ell}$ -vector of the item parameters of the items administered and with  $\mathbf{r}_{\ell}$  the accompanying value of the item indicator variable (see (20)). Without loss of generality we may assume that the total number of distinguished subpopulations is the same as the number of different tests administered:  $T = L$ .

If we use the background information in MML calibration in this case the partition which the observed design variable  $\mathbf{m}_i$  effects is:

$$\left. \begin{aligned} U_{obs,i} &= (\mathbf{X}_{obs,i}, \mathbf{Y}_i) \\ U_{mis,i} &= (\mathbf{X}_{mis,i}, \boldsymbol{\theta}_i) \end{aligned} \right\}, \quad (i = 1, \dots, n). \quad (30)$$

And the distribution of the missing data indicator follows from (23):

$$\mathbf{P}(\mathbf{M}_i = (\mathbf{r}_{\ell}, 1, 0) \mid \mathbf{Y}_i = y^{(\ell)}) = \boldsymbol{\phi}_{\ell}, \quad (\ell = 1, \dots, L). \quad (31)$$

Note that the design vector variable  $\mathbf{M}_i$  has one element more compared to the situations in complete, in random incomplete and in multistage testing indicating the observation of  $\mathbf{Y}_i$ . The  $(k+1)^{th}$  element indicates  $\mathbf{Y}_i$ , the  $(k+2)^{th}$   $\boldsymbol{\theta}_i$ . From (31) it is easily seen that the conditions for ignorability MAR (depending only on observed responses) and D are fulfilled. So the correct likelihood inference can be based on the marginal distribution of the observations. For a randomly sampled student we have:

$$\begin{aligned}
P_{\beta, \gamma}(\mathbf{x}_{obs,i}, \mathbf{Y}_i = y^{(\ell)}) &= \int_{\mathbf{x}_{mis,i}} \int_{\theta_i} P_{\beta, \gamma}(\mathbf{x}_{obs,i}, \mathbf{x}_{mis,i}, \mathbf{Y}_i = y^{(\ell)}, \theta_i) d\theta_i d\mathbf{x}_{mis,i} = \\
\int_{\theta_i} P_{\beta^{(\ell)}}(\mathbf{x}_{obs,i} \mid \mathbf{Y}_i = y^{(\ell)}, \theta_i) \cdot P_{\gamma}(\theta_i \mid \mathbf{Y}_i = y^{(\ell)}) \cdot P(\mathbf{Y}_i = y^{(\ell)}) d\theta_i = \\
\int_{\theta_i} P_{\beta^{(\ell)}}(\mathbf{x}_{obs,i} \mid \theta_i) \cdot g_{\gamma_i}(\theta_i) \cdot \pi_{\ell} d\theta_i = \\
\prod_{\ell=1}^L \pi_{\ell}^{y_{i\ell}} \cdot \prod_{\ell=1}^L \left[ \int_{\theta_i} P_{\beta^{(\ell)}}(\mathbf{x}_{obs,i} \mid \theta_i) \cdot g_{\gamma_i}(\theta_i) d\theta_i \right]^{y_{i\ell}}.
\end{aligned} \tag{32}$$

The likelihood of the total sample is given by:

$$\begin{aligned}
L(\beta, \gamma, \pi; \mathbf{x}_{obs}, \mathbf{y}) &= \prod_{i=1}^n P_{\beta, \gamma}(\mathbf{x}_{obs,i}, \mathbf{Y}_i = \mathbf{y}^{(\ell)}) = \\
\prod_{i=1}^n \prod_{\ell=1}^L \pi_{\ell}^{y_{i\ell}} \cdot \prod_{i=1}^n \prod_{\ell=1}^L \left[ \int_{\theta_i} P_{\beta^{(\ell)}}(\mathbf{x}_{obs,i} \mid \theta_i) \cdot g_{\gamma_i}(\theta_i) d\theta_i \right]^{y_{i\ell}}.
\end{aligned} \tag{33}$$

From (33) it is seen that the likelihood function consist of a term only dependent on the proportions  $\pi_{\ell}$  of the subpopulations in the total population and a term which is a product of  $L$  ordinary marginal likelihood functions. This is because there is always exact one  $\ell$  for which  $y_{i\ell} = 1$ , with the understanding that they not all contain the same item parameters. Standard maximum likelihood estimates  $\hat{\pi}_{\ell}, \ell = 1, \dots, L$  of the proportions can be obtained from the first part. Maximizing the second term with respect to  $\gamma_{\ell}, \ell = 1, \dots, L$  and  $\beta$  will give estimates of  $L$  population parameters and the item parameters. Calibration using the background information in the TTOP case is thus a generalization of standard MML.

If we do not use the background information in the TTOP case, the partition the observed design variable  $\mathbf{m}_i$  establishes becomes:

$$\left. \begin{aligned} U_{obs,i} &= X_{obs,i} \\ U_{mis,i} &= (X_{mis,i}, Y_i, \theta_i) \end{aligned} \right\}, \quad (i = 1, \dots, n). \quad (34)$$

The design distribution is given by:

$$P(\mathbf{M}_i = (\mathbf{r}_\ell, 0, 0) \mid \mathbf{Y}_i = y^{(\ell)}) = \phi_\ell, \quad (\ell = 1, \dots, L). \quad (35)$$

We see that the MAR condition in this case is not fulfilled, because the design distribution depends on values of  $\mathbf{Y}_i$  which are considered as missing if we do not use  $\mathbf{Y}$  in the analyses. Not using  $\mathbf{Y}$  in the TTOP case is not justified by the ignorability principle and can lead to incorrect estimates of the parameters. The next example will illustrate this.

### Example 3

In a simulation study, data were generated according to the following specifications: two non-equivalent samples of 1000 students were drawn from two normal distributions, respectively  $\theta \sim N(-1, 1)$  and  $\theta \sim N(+1, 1)$ . The less able population is administered the first 6 items out of a pool of 9 items. The more able pupils took the last 6 items. So the anchor consisted of 3 items. The responses are generated according to the Rasch model and the item parameters are:  $\beta_1 = -2.0$ ,  $\beta_2 = -1.0$ ,  $\beta_3 = -0.5$ ,  $\beta_4 = \beta_5 = \beta_6 = 0$  and  $\beta_7 = 0.5$ ,  $\beta_8 = 1.0$ ,  $\beta_9 = 2.0$ . So we have a data matrix with the same structure as in a targetted testing design, in which students are assigned to one of the two test booklets on the basis of a background variable. If we estimate the item parameters ignoring the background variable or design variable and apply MML estimation in a standard way with one ability distribution, we get the results given column 3 of Table 4.2.

Table 4.2 Input  $\beta$  and estimated  $\hat{\beta}$  difficulty parameters Rasch model

item	$\beta$ (input)	$\hat{\beta}$ (se); one marginal	$\hat{\beta}$ (se); two marginals
1	-2.0	-1.521 (.080)	-1.979 (.079)
2	-1.0	-0.418 (.072)	-0.938 (.072)
3	-0.5	0.051 (.072)	-0.498 (.073)
4	0	-0.042 (.051)	-0.066 (.053)
5	0	0.032 (.051)	0.014 (.053)
6	0	-0.045 (.051)	-0.069 (.053)
7	0.5	0.046 (.073)	0.589 (.075)
8	1.0	0.417 (.073)	0.952 (.074)
9	2.0	1.480 (.079)	1.996 (.080)
mean		$\hat{\mu} = 0.047$	$\hat{\mu}_1 = -0.986 \quad \hat{\mu}_2 = 1.097$
sd		$\hat{\sigma} = 1.293$	$\hat{\sigma}_1 = 0.954 \quad \hat{\sigma}_2 = 1.129$

We see a clear bias in the estimates of the parameters that were administered in only one of the two non-equivalent samples. The difficulty parameters of the items only administered in the less able group ( $\mathcal{E}\theta = -1.0$ ) are overestimated and are underestimated in the more able group ( $\mathcal{E}\theta = 1.0$ ). If we do not ignore the design variable and estimate with two marginal distributions (33) we get the results in column 4 of Table 4.2, which are seen to be free from systematic bias.  $\square$

In the TTMP situation the background variable is used as a stratification variable: from every subpopulation  $\ell$ ,  $\ell = 1, \dots, L$ , we have a random sample from  $g_{\gamma}(\theta)$  with  $n_{\ell}$  the number of observations in subpopulation  $\ell$  and  $\sum_{\ell=1}^L n_{\ell} = n$  the total sample size. The sampling proportions in the subpopulations,  $\pi_{\ell}^* = n_{\ell}/n$ , can but will in general not be equal to the population proportions  $\pi_{\ell}$ . These population proportions  $\pi_{\ell}$  are not estimable in this case but they must be known in advance. This also means that in the TTMP case the distribution in the total population (29) can only completely be estimated provided the population proportions are known and that we have

samples from every subpopulation,  $n_\ell > 0, \ell = 1, \dots, L$ . Otherwise we are not able to estimate all subpopulation parameters  $\gamma_\ell, \ell = 1, \dots, L$ . Another difference from the TTOP situation is that in TTMP the values of  $\mathbf{Y}$  are known before sampling, so  $\mathbf{Y}$  is not a random variable here. In order to identify the membership of a student of a subpopulation we will have to use the values of  $\mathbf{Y}$ . So we will not consider the simultaneous probability of the observed response vector  $\mathbf{X}_{obs,i}$  and  $\mathbf{Y}_i$  as in the TTOP case (32), but the conditional distribution of  $\mathbf{X}_i$  given  $\mathbf{Y}_i = \mathbf{y}^{(\ell)}$ . The design distribution is given by:

$$P(\mathbf{M}_i = \mathbf{m}_{i_\ell} = (\mathbf{r}_\ell, 0)) = 1 \quad \text{if} \quad \mathbf{Y}_i = \mathbf{y}^{(\ell)}. \quad (36)$$

Compared to (31), the TTOP case  $\mathbf{M}_i$  has one element less, because  $\mathbf{Y}_i$  is not random.

Because of (36) the conditional distribution of a response vector given  $\mathbf{Y}_i = \mathbf{y}^{(\ell)}$  is the same as the conditional distribution of  $\mathbf{X}_{obs,i}$  given the design variable. For a randomly sampled student from subpopulation  $\ell$  we have:

$$\begin{aligned} P_{\beta_{(\ell)}, \gamma_\ell}(\mathbf{x}_{obs,i_\ell} \mid \mathbf{m}_{i_\ell}) &= P_{\beta_{(\ell)}, \gamma_\ell}(\mathbf{x}_{obs,i_\ell} \mid \mathbf{Y}_i = \mathbf{y}^{(\ell)}) = \\ \int_{\mathbf{x}_{mis,i_\ell}} \int_{\theta_{i_\ell}} P_{\beta_{(\ell)}, \gamma_\ell}(\mathbf{x}_{obs,i_\ell}, \mathbf{x}_{mis,i_\ell}, \theta_{i_\ell} \mid \mathbf{Y}_i = \mathbf{y}^{(\ell)}) \cdot P_{\gamma_\ell}(\theta_{i_\ell} \mid \mathbf{Y}_i = \mathbf{y}^{(\ell)}) d\theta_{i_\ell} d\mathbf{x}_{mis,i} &= \\ \int_{\theta_{i_\ell}} P_{\beta_{(\ell)}}(\mathbf{x}_{obs,i_\ell} \mid \theta_{i_\ell}) \cdot g_{\gamma_\ell}(\theta_{i_\ell}) d\theta_{i_\ell}. \end{aligned} \quad (37)$$

And for the total sample we have the likelihood

$$\prod_{\ell=1}^L \prod_{i_\ell=1}^{n_\ell} \int_{\theta_{i_\ell}} P_{\beta_{(\ell)}}(\mathbf{x}_{obs,i_\ell} \mid \theta_{i_\ell}) \cdot g_{\gamma_\ell}(\theta_{i_\ell}) d\theta_{i_\ell}. \quad (38)$$

As before the parameters  $\beta$  and  $\gamma_\ell, \ell=1, \dots, L$  (provided  $n_\ell > 0$ ) can be estimated from (38). It is noted that in the TTMP situation we do not ignore the design variable in the analyses but explicitly condition on it.



If we do not use the background information in the TTMP case this will not lead to correct inferences on the parameters. If we were willing to make the unrealistic extra assumption that all students are randomly drawn from one population with ability distribution  $g_{\gamma^*}(\theta)$  defined by

$$g_{\gamma^*}(\theta) = \sum_{\ell=1}^L \pi_{\ell}^* g_{\gamma_{\ell}}(\theta) = \sum_{\ell=1}^L (n_{\ell}/n) \cdot g_{\gamma_{\ell}}(\theta) \quad (39)$$

then we are in fact in the TTOP situation for which it was shown ((34) and (35)) that by ignoring  $\mathbf{Y}$  the MAR condition for ignorability is not fulfilled.

Summarizing we can say that in MML item calibration in complete testing designs as long as we are sampling from one population there is more or less a free choice of whether the background variable is used in order to get estimates of the item parameters. However when sampling from multiple subpopulations and always in incomplete targeted testing designs, in TTOP as well as TTMP, there is no choice whether the background information  $\mathbf{Y}$  must be used. Not using  $\mathbf{Y}$  never leads to correct inferences on the item parameters or the population parameters. So we are obliged to use the subpopulation structure in MML estimation in order to get a correct estimation procedure. It will also be clear that the parameters of the ability distribution of the total population can only be estimated correctly, even in the case that we have a random sample from one population, via estimating the subpopulation parameters and the population proportions. Although standard computer implementation of MML procedures (e.g., in BILOG-MG, OPLM) have facilities to use  $\mathbf{Y}$ , and always assume one random sample of students, the awareness of the possible problems is not general and in practice many failures are made.

#### ***4.5.2 The conditional model and missing data***

In the preceding section it was shown that in MML estimation in incomplete designs checking Rubins (1976) conditions for ignorability is useful. Only when we are sampling from multiple populations it is not possible to ignore the design

variable (in targeted testing) and explicitly use the design in the analysis. But in all other cases considered checking the standard conditions to be met for ignorability, makes clear whether or not estimating the parameters with MML while ignoring the design variable is justified.

We will elaborate now on whether applying these ignorability checks are also useful in CML estimation. In applying the ignorability principle we fix the random design variable  $\mathbf{M}$  at the observed pattern of missing data  $\mathbf{m}$  and assume that the values  $\mathbf{u}_{obs}$  are realizations of the marginal distribution of  $\mathbf{U}_{obs}$  (11):

$$\int_{\mathbf{u}_{mis}} f(\mathbf{u}_{obs}, \mathbf{u}_{mis}) d\mathbf{u}_{mis} . \quad (40)$$

Remember (15) that the correct distribution of the realizations  $\mathbf{u}_{obs}$ ,

$$\int_{\mathbf{u}_{mis}} f(\mathbf{u}_{obs}, \mathbf{u}_{mis} | \mathbf{m}) d\mathbf{u}_{mis} , \quad (41)$$

the conditional distribution of  $\mathbf{U}_{obs}$  given  $\mathbf{M} = \mathbf{m}$ , is not used in the analysis, but only the marginal distribution of the observed responses. Note that in the CML case, the design variable  $\mathbf{M}_i$  and the item indicator variable  $\mathbf{R}_i$  are the same because the only variables inferred on are the item responses  $\mathbf{X}$ , and  $\boldsymbol{\theta}$  is not treated as a random variable as in MML. It will be clear that ignoring the design variable in CML estimation is only possible if for an individual observed response vector  $\mathbf{X}_{obs,i}$  there exists a sufficient statistic  $S_{obs,i} = S(\mathbf{X}_{obs,i})$  for  $\boldsymbol{\theta}_i$  in the marginal distribution (40). It can easily be shown that in the IRT models we consider, for example in the Rasch model the sum score

$$S_{obs,i} = \sum_{j \in obs} X_{ij} , \quad (42)$$

is not only not sufficient for  $\boldsymbol{\theta}_i$  in the marginal distribution of the observations  $\mathbf{X}_{obs,i}$ , but also not sufficient in the distribution of all observed data  $(\mathbf{X}_{obs,i}, \mathbf{R}_i)$ .  $S_{obs,i}$  is only sufficient in the conditional distribution of the responses given the item indicator variable  $\mathbf{R}_i$ . An example will make this clear.

Assume we have 3 items following the Rasch model with parameters

$\epsilon_i = \exp(-\beta_i)$ ,  $i=1,2,3$  and a random item indicator variable with two possible outcomes ( $0 < \phi < 1$ ):

$$P(\mathbf{R}_i = \mathbf{r}_1 = (1,1,0)) = \phi, \text{ and } P(\mathbf{R}_i = \mathbf{r}_2 = (1,0,1)) = 1 - \phi. \quad (43)$$

In Table 4.3 the relevant probabilities for all outcomes with  $S_{obs} = 1$ , with  $\exp(\theta) = \xi$ , are given.

Table 4.3 Probabilities for all outcomes with  $S_{obs} = 1$ .

$\mathbf{x}_{obs}, \mathbf{r}$	$P(\mathbf{x}_{obs}, \mathbf{r})$	$P(\mathbf{x}_{obs} \mid \mathbf{r}_1)$	$P(\mathbf{x}_{obs} \mid \mathbf{r}_2)$
	(1)	(2)	(3)
10,110	$\frac{\phi \cdot \xi \epsilon_1}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_2)}$	$\frac{\xi \epsilon_1}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_2)}$	0
01,110	$\frac{\phi \cdot \xi \epsilon_2}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_2)}$	$\frac{\xi \epsilon_2}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_2)}$	0
10,101	$\frac{(1 - \phi) \cdot \xi \epsilon_1}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_3)}$	0	$\frac{\xi \epsilon_1}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_3)}$
01,101	$\frac{(1 - \phi) \cdot \xi \epsilon_3}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_3)}$	0	$\frac{\xi \epsilon_3}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_3)}$
1	$\frac{\phi \cdot \xi(\epsilon_1 + \epsilon_2)}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_2)} + \frac{(1 - \phi) \cdot \xi(\epsilon_1 + \epsilon_3)}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_3)}$	$\frac{\xi(\epsilon_1 + \epsilon_2)}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_2)}$	$\frac{\xi(\epsilon_1 + \epsilon_3)}{(1 + \xi \epsilon_1)(1 + \xi \epsilon_3)}$
$S_{obs}$	$P(S_{obs})$	$P(S_{obs} \mid \mathbf{r}_1)$	$P(S_{obs} \mid \mathbf{r}_2)$

Conditioning on  $S_{obs}$  in the joint distribution of  $\mathbf{X}_{obs}$  and  $\mathbf{R}$ , that is, dividing in Table 4.3 the terms in the upper part of column (1) by the term in the lower part, does not cancel the individual parameter  $\xi$ . On the other hand it can easily be checked that in the conditional distributions of  $\mathbf{X}_{obs}$  given  $\mathbf{R}$ ,  $S_{obs}$  is sufficient for  $\xi$ . Divide the upper part terms in column (2) and (3) in Table 4.3 by their

lower part term. In the example the same is easily checked for the outcomes with  $S_{obs}$  is 2 and 0.

In general, the probability of the observed variables can be written as

$$P_{\theta, \beta, \phi}(\mathbf{x}_{obs}, \mathbf{r}) = \prod_i P_{\theta, \beta, \phi}(\mathbf{x}_{obs, i} | \mathbf{r}_i) \cdot P_{\phi}(\mathbf{r}_i) . \quad (44)$$

We use the same notation as before. We distinguish  $T$  values of the design variable  $\mathbf{r}_t, t = 1, \dots, T$ ;  $n_t$  is the number of students taking test  $t$ ;  $\beta_{(t)}$  is the  $k_t$  - vector of the parameters of the items in test  $t$ . We can then rewrite (44) as:

$$P_{\theta, \beta, \phi}(\mathbf{x}_{obs}, \mathbf{r}) = \prod_{t=1}^T \prod_{i=1}^{n_t} P_{\theta, \beta_{(t)}, \phi}(\mathbf{x}_{obs, i} | \mathbf{r}_t) \cdot P_{\phi}(\mathbf{r}_t) . \quad (45)$$

We see in (45) that we have in fact the product of  $T$  complete data likelihoods. For every  $t$  the first factor in the right-hand side of (45) can, as in complete data CML (see (5)), be written as

$$\prod_{i=1}^{n_t} P_{\theta, \beta_{(t)}, \phi}(\mathbf{x}_{obs, i} | \mathbf{r}_t) = \prod_{i=1}^{n_t} P_{\beta_{(t)}}(\mathbf{x}_{obs, i} | s_{obs, i}, \mathbf{r}_t) \cdot P_{\theta, \beta_{(t)}, \phi}(s_{obs, i} | \mathbf{r}_t) . \quad (46)$$

And the first factor in the right-hand side of (46) is again free of any incidental parameters, and

$$L_c = \prod_{t=1}^T \prod_{i=1}^{n_t} P_{\beta_{(t)}}(\mathbf{x}_{obs, i} | s_{obs, i}, \mathbf{r}_t) \quad (47)$$

can be used for CML estimation of  $\beta$ . Note that when estimating the item parameters in this way there are as many different sufficient statistics as there are designs involved.

So we have seen that the standard ignorability checks of Rubin cannot be applied in CML estimation. We have to condition explicitly on the design variable in order to get sufficient statistics for the incidental parameters. But whether it is justified to estimate the item parameters by just maximizing the likelihood (47) depends of course, as in the complete data case, on the properties

of the part of the total likelihood (45) we neglect in that case. The neglected part in CML estimation in incomplete designs is (combining (45), (46) and (47))

$$\prod_{t=1}^T \prod_{i=1}^{n_t} P_{\theta, \beta_{(t)}, \phi}(s_{obs,i}, \mathbf{r}_t) = \prod_{t=1}^T \prod_{i=1}^{n_t} P_{\theta, \beta_{(t)}, \phi}(s_{obs,i} | \mathbf{r}_t) \cdot \prod_{t=1}^T \prod_{i=1}^{n_t} P_{\phi}(\mathbf{r}_t). \quad (48)$$

In (48), the first factor on the right hand side is the product of  $T$  terms, which are also neglected in the complete data case. Because neglecting this part was shown to be possible (Eggen, 2000) without severe consequences, the properties of the marginal distribution of the design variable will be decisive for the justification of neglecting the term. We will discuss the properties of (48) for the three considered design types next.

#### CML in random incomplete designs

In random incomplete designs the design distribution is given by (21). Considering the first factor of the part of the likelihood we neglect in CML (48), we see this factor consists of the product of  $T$  complete data distributions of the sufficient statistics  $\mathbf{S}_{obs}$ , which can be neglected. From the design distribution (21) it is easily seen that the second part of (48),  $P_{\phi}(\mathbf{r}_t)$ , does not depend on the item parameters at all. As a consequence, (48) can be neglected in CML estimation. So CML estimation is justified in random incomplete designs.

#### CML in multistage testing designs

In multistage testing the first part of (48) can be neglected for the same reason as in random incomplete designs. The second part, however, the design distribution in multistage testing designs, is dependent of the observed variables. Given the design distribution (22) we can write the second part as:

$$\prod_{t=1}^T \prod_{i=1}^{n_t} P_{\phi}(\mathbf{R}_i = \mathbf{r}_t) = \prod_{t=1}^T \prod_{i=1}^{n_t} P_{\phi}(\mathbf{R}_i = \mathbf{r}_t | \mathbf{x}_{obs,i}) \cdot P_{\beta_{(obs)}, \theta_i}(\mathbf{x}_{obs,i}). \quad (49)$$

We see that (49) is for every  $t$  directly dependent of the item parameters of the items used for establishing the design. This means that (48), cannot be neglected

in CML estimation. So CML estimation is in this situation not justified because not all random variations in the data relevant for estimating the item parameters are considered in the conditional likelihood. Applying CML in these designs, which is in principle possible in the computer programs for CML, gives incorrect estimates of the item parameters. An example will illustrate this.

*Example 2 (continued).*

The items and the design used are given in example 2. Generated are 4000 responses on these items using a standard normal ability distribution. First the item parameters estimated in the complete design are given in column 3 in Table 4.4. In column 4 of Table 4.4, the results are given of the item parameter estimates in the two stage testing design.

*Table 4.4 CML estimates and standard errors in a two stage testing design*

item	$\beta$ (input)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)
		complete	multistage	multistage
1	0	-0.360 (.033)	-0.360 (.035)	-
2	0	0.004 (.033)	0.060 (.035)	-
3	0	0.024 (.033)	0.028 (.035)	-
4	-1.25	-1.284 (.037)	-1.709 (.049)	-1.326 (.053)
5	-1.0	-0.990 (.036)	-1.419 (.048)	-1.021 (.052)
6	-0.5	-0.445 (.034)	-0.467 (.035)	-0.452 (.035)
7	0.5	0.506 (.034)	0.535 (.035)	0.517 (.036)
8	1.0	0.964 (.035)	1.387 (.047)	0.989 (.051)
9	1.25	1.257 (.037)	1.674 (.048)	1.293 (.052)

It is clear that applying CML estimation in this two stage testing design gives systematic errors in the item parameter estimates: the item parameters of the easy items (4 and 5) are underestimated, and the parameters of the hard items are overestimated.

The last column of Table 4.4 gives the results in case the item parameters of the routing test are not estimated themselves. It is seen that in that case CML gives correct estimates on the other items. This can be understood by the fact that distribution of the design variable (46) is not dependent on the parameters to be estimated. If we denote the indices of the observed items in the routing test with  $obs1$  and the parameter vector with  $\beta^{(1)}$ , and the other with  $obs2$  and  $\beta^{(2)}$  then in CML estimation of the items that are not in the routing test the following likelihood is used:

$$L_c = \prod_{t=1}^T \prod_{i=1}^{n_t} P_{\beta_{(t)}^{(2)}}(\mathbf{x}_{obs2,i} | s_{obs2,i}, \mathbf{r}_t).$$

And the distribution of the design which is neglected in the estimation is given by

$$\prod_{t=1}^T \prod_{i=1}^{n_t} P_{\phi}(\mathbf{R}_i = \mathbf{r}_t | \mathbf{x}_{obs1,i}) \cdot P_{\beta_{(obs1)}, \theta_i}(\mathbf{x}_{obs1,i}).$$

does not depend on the parameters  $\beta^{(2)}$ , which are estimated.  $\square$

Following the procedure given in Example 2 is a possible practical solution if the items are to be estimated with CML a two stage testing design. Glas (1988) showed that another possible approach for CML in multistage testing, conditioning on the scores for every stage of the design, fails, because it results in separate calibrations for the items in a stage, which can not be connected on the same scale.

### CML in targeted testing designs

In targeted testing designs the value of a background variable  $\mathbf{Y}$  determines the design. The design distribution is given by (23). Before we made the distinction between the two sampling roles  $\mathbf{Y}$  can play in the design and using or not using  $\mathbf{Y}$  was of utmost importance in MML estimation. In CML estimation, however, these distinctions are not relevant.

Firstly, consider complete testing designs in the presence of background information. The simultaneous probability of the response vector  $\mathbf{X}_i$  and of  $\mathbf{Y}_i$  of student  $i$  is given by

$$P_{\theta, \beta, \pi_q}(\mathbf{x}_i, \mathbf{Y}_i = \mathbf{y}^{(\ell)}) = P_{\theta, \beta}(\mathbf{x}_i \mid \mathbf{Y}_i = \mathbf{y}^{(\ell)}) \cdot P_{\pi_q}(\mathbf{Y}_i = \mathbf{y}^{(\ell)}) . \quad (50)$$

Conditioning on the sufficient statistic  $\mathbf{S}_i$  gives:

$$\begin{aligned} P_{\theta, \beta, \pi_q}(\mathbf{x}_i, \mathbf{Y}_i = \mathbf{y}^{(\ell)}) &= P_{\theta, \beta}(\mathbf{x}_i \mid \mathbf{S}_i, \mathbf{Y}_i = \mathbf{y}^{(\ell)}) \cdot P_{\theta, \beta}(\mathbf{S}_i \mid \mathbf{Y}_i = \mathbf{y}^{(\ell)}) \cdot P_{\pi_q}(\mathbf{Y}_i = \mathbf{y}^{(\ell)}) \\ &= P_{\beta}(\mathbf{x}_i \mid \mathbf{S}_i) \cdot P_{\theta, \beta}(\mathbf{S}_i \mid \mathbf{Y}_i = \mathbf{y}^{(\ell)}) \cdot P_{\pi_q}(\mathbf{Y}_i = \mathbf{y}^{(\ell)}) . \end{aligned} \quad (51)$$

In (51)  $\mathbf{Y}_i = \mathbf{y}^{(\ell)}$  cancels in  $P_{\beta}(\mathbf{x}_i \mid \mathbf{S}_i)$  because given  $\theta_i$  the item responses are not dependent of any other characteristic of the students (2). The complete likelihood of the sample is given by:

$$\prod_i P_{\beta}(\mathbf{x}_i \mid \mathbf{S}_i) \cdot \prod_i P_{\theta, \beta}(\mathbf{S}_i \mid \mathbf{Y}_i = \mathbf{y}^{(\ell)}) \cdot P_{\pi_q}(\mathbf{Y}_i = \mathbf{y}^{(\ell)}) . \quad (52)$$

From (52), the first factor is used in CML estimation. And, as before the second factor is always discarded in CML estimation and the third factor is independent of it. So CML is a justified procedure to estimate  $\beta$ . Furthermore it is clear that the background information is in fact always used in the analyses, since it defines the design, but it appears only in that part of the likelihood which can be neglected in CML estimation. If we would have samples from multiple populations all the above still holds. The only change we have to make is that we start with  $P_{\theta, \beta}(\mathbf{x}_i \mid \mathbf{Y}_i = \mathbf{y}^{(\ell)})$  with as a consequence that  $P_{\pi_q}(\mathbf{Y}_i = \mathbf{y}^{(\ell)})$  cancels in (51) and (52). So it can be concluded that in CML estimation all the sample information is in that part of the total likelihood which is justified to be neglected. The independence of CML estimation of the actual sample available for estimation can be understood in this way.



Next, we consider incomplete targeted testing. Here we distinguish as many values ( $L$ ) of the design variable  $\mathbf{r}_i$  as we distinguish values of the background variable  $\mathbf{Y}_i$ . If we rewrite the total likelihood as before ((45), (47) and (48)) we see that the conditional likelihood to be maximized is:

$$\prod_{\ell=1}^L \prod_{i=1}^{n_{\ell}} P_{\beta_{(\ell)}}(\mathbf{x}_{obs,i} | s_{obs,i}, \mathbf{r}_{\ell}, \mathbf{Y}_i = y^{(\ell)}), \quad (53)$$

and the neglected part becomes

$$\prod_{\ell=1}^L \prod_{i=1}^{n_{\ell}} P_{\theta, \beta_{(\ell)}, \phi, \pi}(s_{obs,i}, \mathbf{r}_{\ell}, \mathbf{Y}_i = y^{(\ell)}) = \quad (54)$$

$$\prod_{\ell=1}^L \prod_{i=1}^{n_{\ell}} P_{\theta, \beta_{(\ell)}, \phi, \pi}(s_{obs,i} | \mathbf{r}_{\ell}, \mathbf{Y}_i = y^{(\ell)}) \cdot \prod_{\ell=1}^L \prod_{i=1}^{n_{\ell}} P_{\phi, \pi}(\mathbf{r}_{\ell}, \mathbf{Y}_i = y^{(\ell)}).$$

From the design distribution (23) it is seen that the second part of the right hand side of (54) is independent of the item parameters which are to be estimated. So CML estimation, on the basis of the conditional likelihood (53), is justified in targeted testing.

### Example 3 (continued)

If we estimate the item parameters of example 3 with CML, we see in results of Table 4.5. that targeted testing does not cause any systematic errors in the item parameter estimates.

Table 4.5: Input  $\beta$  and estimated  $\hat{\beta}$  difficulty parameters Rasch model

item	$\beta$ (input)	$\hat{\beta}$ (se);CML
1	-2.0	-1.980 (.080)
2	-1.0	-0.935 (.072)
3	-0.5	0.497 (.073)
4	0	-0.066 (.053)
5	0	0.015 (.053)
6	0	-0.069 (.053)
7	0.5	0.592 (.075)
8	1.0	0.954 (.074)
9	2.0	1.986 (.080)

## 4.6 Conclusion

In this chapter it was shown for the three most common stochastic design types under which conditions item calibration is possible. It was seen that in MML estimation Rubins ignorability can directly be applied to justify the missing data procedures. In CML estimation this was seen not to be the case. In CML the design is never ignored and must always be an explicit part of the conditional likelihood. In CML we in fact always work with the combination of as many complete data likelihoods as there are designs. The key condition for justifying CML is in the dependence of the distribution of the design variable on the item parameters which are to be estimated.

Summarized it can be said that in random incomplete designs both MML and CML are possible. In multistage testing designs MML is always a good option for item calibration. CML estimation is in multistage testing in general not justified. It was shown, that in a two testing design a practical feasible solution is, to conduct the CML estimation without estimating the item parameters of the routing test. In targeted testing CML is always possible. MML estimation gives sometimes problems. If one knows, for instance by stratified sampling, that in the testing design the assignments to the test booklets is according to these strata,

MML estimation is justified when as many marginal ability distributions are specified as strata or designs. Ignoring the background variable gives biased results.

It was noticed that because standard computer algorithms for MML assume a random sample from one population in practice many failures are made when we have in fact not one random but a stratified sample or when we have a targeted testing design. In CML computer algorithms data from multistage testing designs can give incorrect results.

It should be noticed that all the principles elaborated for the three basic designs can also be applied in combination, when we have designs in which properties of the basic designs are combined.

Finally it is remarked that in this chapter all results are for convenience illustrated by the simple one-parameter logistic model for dichotomously scored items. But all results also apply, whenever CML or MML is applicable, for models for polytomously scored items and for models with more than one item parameter.

## 4.7 References

- Andersen, E.B. (1973). *Conditional inference and models for measuring*. Unpublished Ph.D.Thesis, Copenhagen: Mentalhygiejnisk Forlag.
- Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Eggen, T.J.H.M. (1990). Innovative procedures in the calibration of measurement scales. In W.H. Schreiber & K. Ingenkamp: *International developments in large-scale assessment* (pp 199-212). Windsor, Berksley: NFER-NELSON.
- Eggen, T.J.H.M. (2000). On the loss of information in conditional maximum likelihood estimation. *Psychometrika*, 65, 337-362.
- Eggen, T.J.H.M. & Verhelst, N.D. *Item calibration in incomplete testing designs*. Measurement and Research Department Reports 92-3. Arnhem: Cito.
- Fischer, G.H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika*, 46, 59-77.
- Glas, C.A.W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics*, 13, 45-52.
- Glas, C.A.W. (1989). *Contributions to estimating and testing Rasch models*. Unpublished Ph.D. Thesis, Arnhem: Cito.
- Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 887-903.
- Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
- Lord, F.M. (1971). A theoretical study of two-stage testing. *Psychometrika*, 36, 227-242.
- Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum Associates.
- Mislevy, R.J. & Wu, P-K (1988). Inferring examinee ability when some item responses are missing. *Research Report RR-88-48-ONR*. Princeton: Educational Testing Service.

- Mislevy, R.J. & Sheenan, K.M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika*, 54, 661-680.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). One-parameter logistic model (OPLM). [Computer software]. Arnhem: Cito.
- Verhelst, N.D. & Eggen, T.J.H.M. (1989). Psychometrische en statistische aspecten van peilingsonderzoek. [Psychometric and statistical aspects of national assessment research.] *PPON-rapport nr.4*. Arnhem: Cito.
- Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). BILOG-MG [Computer Software]. Chicago: Scientific Software International.

## Chapter 5

### Computerized adaptive testing for classifying examinees into three categories<sup>1</sup>

---

<sup>1</sup>This chapter is a minor revised reprint of: Eggen, T.J.H.M. & Straetmans, G.J.J.M. (2000). Computerized adaptive testing for classifying examinees in three categories. *Educational and Psychological Measurement*, 60, 713-734

## **Abstract**

The objective of this study was to explore the possibilities for using computerized adaptive testing in situations in which examinees are to be classified into one of three categories. Testing algorithms with two different statistical computation procedures are described and evaluated. The first computation procedure is based on statistical testing (the sequential probability ratio test) and the other on statistical estimation (weighted maximum likelihood). Combined with the computation procedures, item selection methods based on maximum information (MI) considering content and exposure control are studied. The measurement quality of the proposed testing algorithms is reported on the basis of simulation studies using an IRT-calibrated item bank involving mathematics. The main results of the study are that a reduction of at least 22% in the mean number of required items can be expected in a computerized adaptive test (CAT) compared to an existing paper-and-pencil placement test. Furthermore, for the three-way classification problem, statistical testing is a promising alternative to statistical estimation. Finally, it is concluded that imposing constraints on the MI selection strategy in the form of content and/or exposure control does not negatively affect the quality of the testing algorithms.

## **5.1 Introduction**

Computerized adaptive tests (CATs) were originally developed to obtain an efficient estimate of an examinee's ability. However, CATs have also shown to be useful in classification problems. Weiss and Kingsbury (1984), Lewis and Sheenan (1990), and Spray and Reckase (1994) described CATs for situations in which the main interest is not in estimating the ability of an examinee, but in classifying the examinee into one of two categories (e.g., pass/fail or master/non-master). The purpose of this chapter is to explore the possibilities for computerized adaptive testing based on item response theory (IRT) in a situation in which examinees are to be classified into one of three categories.

The core of a CAT is the testing algorithm. The algorithm consists of two main components. The first part is a statistical computation procedure which infers the ability of the examinee on the basis of responses to items. The second part is an item selection method, which after the administration of every item, adapts the composition of the test to the ability demonstrated up to that point. Using a calibrated item bank, the algorithm controls the start, the continuation, and the termination of a CAT. Item administration is continued until the ability of an examinee, or a decision to be taken on it, can be reported with a specified level of accuracy.

In IRT-based CATs for classification problems with two categories, two approaches in the statistical computation procedure are currently being used. The first approach is based on statistical estimation (Weiss & Kingsbury, 1984), in which after the administration of each item a confidence interval is constructed. Another item is administered as long as this interval contains a predetermined cutting score, and after the interval no longer contains the cut score one of the two decisions is taken. The second approach uses statistical testing, applying the sequential probability ratio test (SPRT) (Wald, 1947). The first applications of the SPRT for the two-way classification problem assumed that the probability of a correct answer was the same for all items. Reckase (1983) was the first to apply



a modification of the SPRT for IRT-based adaptive testing, which allowed for different probabilities of correct answers to items.

In adaptive tests in which the main purpose is to estimate the ability of an examinee, the item selection method that maximizes the item information at the current ability estimate is in common use (Thissen & Mislevy, 2000). This item selection method can also be applied in the two-way classification problem. However, Spray and Reckase (1984) reported that the item selection method that maximizes the item information at the cutting point of the classification, is (a bit) better than selecting items which have maximum information at the current estimate of an examinee. Their conclusion applies to the statistical estimation as well as the statistical testing approach in the two-way classification problem.

For both approaches it has been shown that they are more efficient than conventional tests in which a common set of items is administered to every examinee (Kingsbury & Weiss, 1983). However, a preference for one of the two approaches has not yet been established. In two studies, attempts were made to compare the results of both approaches. Kingsbury and Weiss (1983) reported that both approaches are greatly affected by the characteristics of the available items. Also, they reported different results depending on the item response model used. In some cases either the testing or the estimation approach may be preferred. It must be noted that in their comparisons the item selection was based on maximum information at the current estimate. Recently, Spray and Reckase (1996) also compared the two approaches. In their simulation studies, they used the item selection method with maximum information at the cutting point and considered only one item response model. In their conclusion they reported a preference for the testing approach, but limited their conclusion explicitly to the situation they studied.

In the present chapter, the possibilities for using computerized adaptive testing in a situation in which examinees are to be classified into one of three categories are explored. Generalizing from the two-way classification problem, two statistical computation procedures are described and evaluated. These computation procedures, one based on statistical estimation and the other on

statistical testing, will be evaluated in combination with two item selection methods. In the first method the next item to be administered is the one which has maximum information (MI) at the current ability estimate of the examinee. In the second method, following Spray and Reckase (1994), items which have MI at a cutting point of the classification are selected. Although MI item selection in CATs provides optimal measurement characteristics, additional factors are considered in practice (Wainer, 2000). In this study the effects of adding content control and exposure control to the MI item selection methods for the three-way classification problem were also investigated.

## **5.2 Context**

The goal of adult basic education in the Netherlands is to provide every adult with the knowledge and abilities that are indispensable for functioning both as an individual and as a member of society. One of the courses provided in this program is a mathematics course that is offered at three different levels of difficulty. A placement test is used to assign prospective students to one of these three course levels. As the ability of the students varies considerably, the placement test being used is a two-stage test (Lord, 1971). In the first stage, all examinees take a routing test of 15 items, which has a difficulty level targeted at the average ability of the prospective students. After the routing test, the examinees, depending on the results on the routing test, take one of three measurement tests of 10 items each, which differ in difficulty.

There are certain drawbacks to the current paper-and-pencil placement test: the test administration procedure is complicated, it is not possible to ensure item confidentiality, and the measurement accuracy for large groups of examinees is limited, especially for examinees at extreme ability levels. Replacing the paper-and-pencil test by a CAT is considered one way to overcome these problems.

### 5.2.1 The mathematics item bank

For the adaptive test, an IRT-calibrated item bank is available. The items in the bank belong to one of three content subdomains: mental arithmetic/estimating (A), measuring/geometry (B), and the other elements of the curriculum (C). Despite these three subdomains, the items were shown to fit a one-dimensional IRT model. The basic equation of the model used (Verhelst & Glas, 1995) was

$$p_i(\theta) = P(X_i = 1 | \theta) = \frac{\exp a_i(\theta - b_i)}{1 + \exp a_i(\theta - b_i)}. \quad (1)$$

The response to an item  $x_i$  is either correct (1) or incorrect (0). The probability of a correct response increases with the latent ability  $\theta$  and depends on two item characteristics: the difficulty parameter,  $b_i$ , and the discrimination index,  $a_i$ .

In the calibration study, 268 items were administered to a sample of 1198 students in an incomplete design in which each student was administered one of 16 different, though overlapping, booklets with about 43 items. Using the OPLM computer program (Verhelst, Glas & Verstralen, 1995), the scaling resulted in an item bank with 250 items (48, 49, and 153 belonging to the subdomains A, B, and C, respectively). The scale was established by imposing the constraints that for the mean item difficulty  $\sum \hat{b}_i = 0$  and that the discrimination indices are positive integers. The geometric mean of the discrimination indices was:  $(\prod a_i)^{1/250} = 3.09$ . The distribution of the ability  $\theta$  in the population was estimated to be normal with a mean of .294 and a standard deviation of .522. The scaling result is summarized in Figure 5.1.

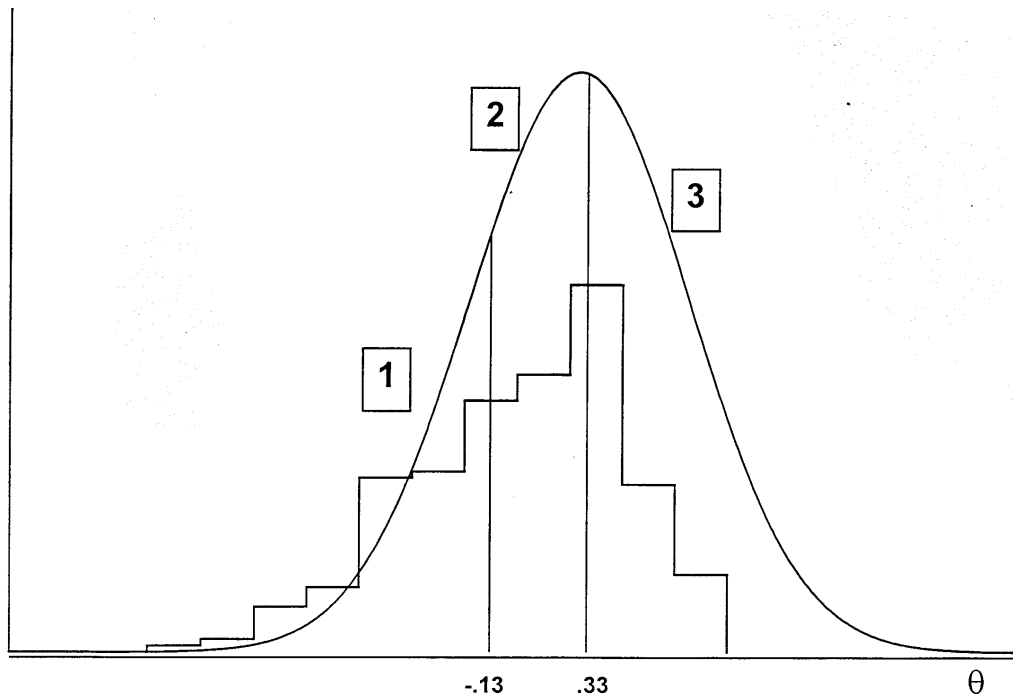


Figure 5.1. Item bank mathematics; histogram difficulties, distribution ability

The cutting points for classification of the examinees into one of the three levels are also shown in Figure 5.1. These were determined using the data from the calibration study and judgments of subject specialists. The procedure followed was:

- (1) subject matter specialists defined subsets of items in the item bank by labeling them as level 1, 2, or 3 items;
- (2) for each subset, using the IRT model (1) the expected score on this subset in the population was computed;
- (3) the lower cutting point was then defined as the ability value at which the expected score is 70% of the maximum score on all the level 1 items; the higher cutting point was defined similarly by calculating the ability value at which the expected score is 70% of the maximum score on the level 2 items.

The resulting cutting point between level 1 and 2 was  $\theta_1 = -.13$  and that between level 2 and 3 was  $\theta_2 = .33$ . A rough inspection of the item bank showed that there seems to be a satisfactory spread of the difficulties of the items on the latent

ability scale. Although the bank contained relatively more easier items, the difficulties of a fair number of items were concentrated near the cutting points.

### **5.3 Research questions**

The main research question in this study was: which testing algorithm is most suitable for the computerized adaptive placement test for mathematics, given a number of practical requirements? From the measurement point of view, evidence was sought to justify the replacement of the current paper-and-pencil placement test by a CAT. The practical requirements were that a CAT may not exceed 25 items and that there should be possibilities for controlling the content of a CAT and the exposure rates of the items. Three additional more specific questions were also examined. Which statistical computation procedures are suitable for classifying examinees into one of three different levels? Which item selection methods should be considered? How do the testing algorithms operate in terms of measurement accuracy, the number of misclassifications, measurement efficiency, adherence to content specifications, and the distribution of exposure rates over the item bank?

### **5.4 Statistical computation procedures**

In a testing algorithm, the statistical computation procedure leads to a decision on the examinee based on the item responses. The inference is made by considering the likelihood function of the examinee's ability  $\theta$ . Given the scores on  $k$  items ( $x_i, i = 1, \dots, k$ ) and the parameters of the items, this function is

$$L_k(\theta; x_1, \dots, x_k) = L_k(\theta; \underline{x}) = \prod_{i=1}^k p_i(\theta)^{x_i} (1 - p_i(\theta))^{1-x_i}. \quad (2)$$

In (2), the IRT model, (1), is substituted.

After each item, it is determined whether another item should be administered or whether testing should be stopped and a decision on the examinee be made. There are two statistical approaches to deal with in the classification problem: statistical estimation and statistical testing. Both approaches will be described briefly.

### 5.4.1 Statistical estimation in the testing algorithm

Because of its good statistical properties, the weighted maximum likelihood (WML) method proposed by Warm (1989) is used for estimating the ability. Based on the scores on  $k$  items, this estimate and its standard error follow from an iterative maximization procedure:

$$\hat{\theta}_k = \max_{\theta} \left( \sum_{i=1}^k I_i(\theta) \right)^{1/2} \cdot \prod_{i=1}^k p_i(\theta)^{x_i} (1 - p_i(\theta))^{1-x_i}. \quad (3)$$

The function which is maximized has two parts. The second part is the likelihood function of the ability (2); the first part is the weight attributed to this likelihood function. This expression contains the item information function  $I_i(\theta)$ . In the IRT model (1), the item information function is given by

$$I_i(\theta) = a_i^2 p_i(\theta)(1 - p_i(\theta)). \quad (4)$$

To classify the examinees, the procedure proposed by Weiss and Kingsbury (1984) is used. After  $k$  items, an estimate is made of the examinee's ability  $\hat{\theta}_k$  and of its standard error  $se(\hat{\theta}_k)$ . Next, a confidence interval  $(\hat{\theta}_k - \gamma \cdot se(\hat{\theta}_k), \hat{\theta}_k + \gamma \cdot se(\hat{\theta}_k))$  for the examinee's true ability  $\theta$  is constructed.  $\gamma$  is a constant, determined by the required accuracy. The algorithm delivers another item as long as there is a cutting point,  $\theta_1$  or  $\theta_2$ , within the interval; if not, a decision is made according to the decision rules set out in Table 5.1.

Table 5.1. Decision rules of adaptive test with statistical estimation

if	Decision
$\hat{\theta}_k + \gamma \cdot se(\hat{\theta}_k) < - .13$	level 1
$\hat{\theta}_k - \gamma \cdot se(\hat{\theta}_k) > - .13$ and $\hat{\theta}_k + \gamma \cdot se(\hat{\theta}_k) < .33$	level 2
$\hat{\theta}_k - \gamma \cdot se(\hat{\theta}_k) > .33$	level 3
else	continue testing

Weiss and Kingsbury (1984) have shown that the procedure for classifying examinees is more efficient and has higher classification validity than a conventional test. This procedure proposed here differs from the one used by Weiss and Kingsbury (1984) in that a non-Bayesian estimation method is used.

### 5.4.2 Statistical testing in the testing algorithm

As an alternative to the estimation procedure, a statistical testing procedure can be used to solve the classification problem. Proposed is a generalization of a procedure used earlier by Reckase (1983), based on the Sequential Probability Ratio Test (SPRT) (Wald, 1947).

First, so-called indifference zones,  $\delta_{..}$ , are defined. These are small areas around the cutting points  $\theta_1$  and  $\theta_2$  in which, owing to measurement fallibility, making the right decision can never be assured. After formulating the statistical hypotheses, the acceptable probabilities of incorrect decisions or decision error rates must be specified. Figure 5.2 presents the problem schematically.

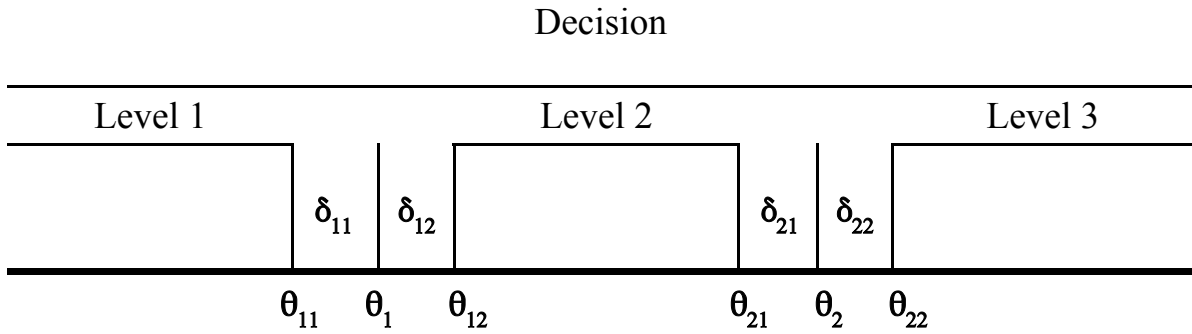


Figure 5.2 Schematic representation of the classification problem with three categories.

The hypotheses are:

H0\_1:  $\theta \leq \theta_{11}$  (level 1)

H1\_1:  $\theta \geq \theta_{12}$  (higher than 1);

H0\_2:  $\theta \leq \theta_{21}$  (lower than 3)

H1\_2:  $\theta \geq \theta_{22}$  (level 3).

The acceptable decision error rates are, with  $\alpha_1$ ,  $\alpha_2$ ,  $\beta_1$ , and  $\beta_2$  small constants, specified as follows:

$P(\text{accept H0\_1} | \text{H0\_1 is true}) \geq 1 - \alpha_1$     $P(\text{accept H0\_1} | \text{H1\_1 is true}) \leq \beta_1$

$P(\text{accept H0\_2} | \text{H0\_2 is true}) \geq 1 - \alpha_2$     $P(\text{accept H0\_2} | \text{H1\_2 is true}) \leq \beta_2$

For each pair of hypotheses (H0\_1 against H1\_1; H0\_2 against H1\_2) the test, meeting the decision error rates, can be carried out using the SPRT (Wald, 1947). The test statistic used is the ratio between the values of the likelihood function (2) under the alternative hypothesis and the null hypothesis:

$$LR_k(\theta_{12}, \theta_{11}; \underline{x}) = \frac{L_k(\theta_{12}; \underline{x})}{L_k(\theta_{11}; \underline{x})}.$$

The test for H0\_1 against H1\_1 operates as follows:

Continue sampling if  $\beta_1/(1 - \alpha_1) < LR_k(\theta_{12}, \theta_{11}; \underline{x}) < (1 - \beta_1)/\alpha_1$  (the critical inequality of the test); accept H0\_1 (level 1) if:  $LR_k(\theta_{12}, \theta_{11}; \underline{x}) \leq \beta_1/(1 - \alpha_1)$ ; and reject H0\_1 (level 2 or 3) if:  $LR_k(\theta_{12}, \theta_{11}; \underline{x}) \geq (1 - \beta_1)/\alpha_1$ .

To solve the three-way classification problem the two SPRTs can be combined, as represented in Table 5.2, to take unequivocal decisions.

Table 5.2. Decisions based on combination of two SPRTs

Decision Test 2	Decision Test 1	
	1	2 or 3
level 1 or 2	1	2
level 3	impossible	3

This generalization of the SPRT is known as Sobel and Wald's (1949) combination procedure. It can easily be shown that, by using the IRT model (1), the impossible solution indeed never occurs and that the critical inequality for the test for H0\_1 against H1\_1 can be written as follows:

$$\frac{\ln \beta_1 / (1 - \alpha_1) - C_{k\theta_{11}\theta_{12}}}{\theta_{12} - \theta_{11}} < \sum_{i=1}^k a_i x_i < \frac{\ln (1 - \beta_1) / \alpha_1 - C_{k\theta_{11}\theta_{12}}}{\theta_{12} - \theta_{11}} \quad (5)$$

$$\text{And in (5) } C_{k\theta_{11}\theta_{12}} = \sum_{i=1}^k \ln \frac{1 + \exp a_i(\theta_{11} - b_i)}{1 + \exp a_i(\theta_{12} - b_i)} = \sum_{i=1}^k \ln \frac{1 - p_i(\theta_{12})}{1 - p_i(\theta_{11})} = \sum_{i=1}^k \ln \frac{q_i(\theta_{12})}{q_i(\theta_{11})}$$

which only depends on the item parameters and on constants in the statistical testing procedure,  $\theta_{11}$  and  $\theta_{12}$ , that are chosen beforehand. The evaluation of the critical inequality is quite easy because it involves only the observed weighted



score and known constants. Table 5.3 represents the decision rules based on the double SPRT in which, for the sake of convenience, it is assumed that  $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = \alpha$ ,  $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = \delta$  and  $\ln(1 - \alpha)/\alpha = A$ .

Table 5.3. Decision rules of adaptive test using statistical testing

if	Decision
$\sum_{i=1}^k a_i x_i \leq \frac{-A - \sum_{i=1}^k \ln \left( \frac{q_i(-.13 + \delta)}{q_i(-.13 - \delta)} \right)}{2\delta}$	level 1
$\frac{A - \sum_{i=1}^k \ln \left( \frac{q_i(-.13 + \delta)}{q_i(-.13 - \delta)} \right)}{2\delta} \leq \sum_{i=1}^k a_i x_i \leq \frac{-A - \sum_{i=1}^k \ln \left( \frac{q_i(.33 + \delta)}{q_i(.33 - \delta)} \right)}{2\delta}$	level 2
$\sum_{i=1}^k a_i x_i \geq \frac{A - \sum_{i=1}^k \ln \left( \frac{q_i(.33 + \delta)}{q_i(.33 - \delta)} \right)}{2\delta}$	level 3
else	continue testing

It is noted that Spray (1993) independently proposed also extensions of the use of the SPRT for classification in three and even  $k$  categories. Her generalization is based on the combination procedure developed by Armitage (1950), which uses the simultaneous application of three SPRTs for classification into three categories, instead of the only two needed in the Sobel and Wald (1949) combination procedure proposed here. It is beyond the scope of this chapter to investigate in detail the properties of these two combination procedures of SPRTs.

## **5.5 Item selection**

In the testing algorithm, the item selection procedure chooses items from the item bank that are adapted to the performance of the examinee as determined by the computation procedure. A special position is taken by the starting procedure because it is assumed that, before the first test administration, the examinee's ability is completely unknown. The starting procedure and the series of selection procedures that are to be investigated for possible implementation in the mathematics placement test are described in this section. Selection procedures that select testlets instead of items, applied by Lewis and Sheenan (1990) among others, will not be considered.

### ***5.5.1 Starting procedure***

The starting procedure for the mathematics placement test operates as follows. Fifty-four relatively easy items are selected from the item bank of 250 items. An examinee is presented one randomly chosen relatively easy item from each of the three content subdomains. There are three reasons for using this starting procedure. The most important is that the target population of examinees partly consists of people who do not feel confident working with a computer. Easy items at the beginning help them overcome their fear of the test and the computer. The second reason is that it is hardly possible to make an optimal choice of items in accordance with the examinee's ability after only one or two items because the first estimates of the ability will inevitably be very inaccurate. Third, the starting procedure, drawing upon the three different subdomains, contributes to the content validity of the test for the domain of mathematics.

### ***5.5.2 Item selection procedures***

In connection with the computation procedures, five item selection procedures have been investigated: (1) random (R); (2) maximum information (MI); (3) maximum information with content control (MI+C); (4) maximum information

with exposure control (MI+E); and (5) maximum information with both content control and exposure control (MI+C+E). The first procedure randomly selects the

next item from the available item bank, excluding items used before. The other procedures select an item for an examinee in such a way that the item information (4) is maximal for that particular examinee. ‘Maximum information’ in this context can have one of the three following meanings.

In the case of statistical estimation as computation procedure, there are two issues:

- (a) The next item selected is the item for which the information at the current ability estimate (CE=current estimate) is maximal. Select the item  $i$  for which:  $\max_i I_i(\hat{\theta}_k)$ . This item selection method is generally accepted as the best non-bayesian in case adaptive testing is used for estimating the ability of an examinee (Thissen & Mislevy, 2000).
- (b) Spray and Reckase (1994) suggested that, with regard to classification problems involving one cutting point, it is more efficient (shorter average test length) to select items that have MI at that cutting point rather than at the current ability estimate. The corresponding selection method is as follows: select an item with MI at the cutting point nearest to the current ability estimate; the minimum is determined of  $|- .13 - \hat{\theta}_k|$  and  $|.33 - \hat{\theta}_k|$ . This option is indicated as NC (nearest cutting point).

If statistical testing is being used as the computation procedure, no ability estimates are made and a variation of (b) is used instead.

- (c) First, the midpoints of the critical inequality intervals of the statistical tests are computed. (In (5) the critical inequality for the test around the cutting point  $\theta_1 = -.13$  is given). The midpoint for the test around  $\theta_1 = -.13$  is  $- C_{k\theta_{11}\theta_{12}} / \delta$ ; and for the test around  $\theta_2 = .33$ :  $- C_{k\theta_{21}\theta_{22}} / \delta$ . Next, it is determined to which midpoint the current weighted score of the examinee is closest: the minimum of  $|\sum_i a_i x_i + C_{k\theta_{11}\theta_{12}} / \delta|$  and

$|\sum_i a_i x_i + C_{k\theta_{21}\theta_{22}} / \delta|$  is determined. The item is selected which has MI at the cutting point corresponding to the midpoint of the critical inequality interval that was found to be closest to the examinee's score.

Maximum information is a psychometric criterion for item selection. However, the practical requirements of test composition can often only be met by constraining this psychometric criterion. Two such constraints were investigated.

The first is related to content control: the adaptive test should meet certain content specifications. In the case of the mathematics placement test the content control took the following form: the preliminary specification was that 16% of the items on a test would be from subdomain A, 20% from subdomain B, and 64% from subdomain C. In order to achieve this, the Kingsbury and Zara (1989, 1991) approach was followed. After each administered item, the difference between the desired and achieved percentage of items selected from each subdomain was determined. The next step was that the item with MI from the domain, for which this difference was largest, was selected.

The second constraint investigated has to do with exposure control. In practice, it often occurs that some items from the available item bank are used very frequently while others are hardly used at all. The purpose of controlling the exposure rates is to avoid possible problems of overexposure and underutilization of items. A simplified form of the Sympton and Hetter (1985) method was used in the placement test of mathematics. When an item has been selected, a random number  $g$  is drawn from the interval (0,1). If  $g > .5$ , the item is administered, if not the procedure is continued by selecting the next most informative item. Items that have been rejected once by this control cannot be selected again for a particular examinee.

## 5.6 Design of the simulation studies

The performance of the computation procedures and item selection methods in the placement test was investigated by means of simulation studies. The first part of the simulation study was concerned only with testing algorithms which use

statistical estimation as computation procedure. The different item selection methods were compared on the measurement inaccuracy after  $k$  items; that is, the mean absolute difference between true and estimated ability:

$$IA_k(\theta) = \frac{1}{N} \sum_{v=1}^N |\theta - \hat{\theta}_{k,v}|. \quad (6)$$

In the main part of the simulation study, the performance of the testing algorithms in the conditions of the placement test were evaluated. The mean number of items required to make a decision,  $k$ , and the classification accuracy, the percentages of correct decisions, %, are reported. In these simulations, the preset accuracy of the testing algorithms was varied. In the statistical estimation computation procedure, two levels of accuracy are reported:  $\gamma$  is 1.034 and 1.644, corresponding to confidence intervals of 70% and 90% respectively. In the statistical testing computation procedure, the acceptable decision error rates varied:  $\alpha$  is .05, .075 and .1. In addition, the indifference zone was varied:  $\delta$  is .1 and .1333 at  $\alpha=.075$ . In both the statistical estimation and the statistical testing, these stopping rules were constrained by a maximum test length:  $k_{\max} = 25$ . If this maximum number of items was needed, a deviation was made from the decision rules in Tables 5.2 and 5.3 insofar that the most obvious decision was taken. The decision rules at  $k_{\max} = 25$  are given in Table 5.4.

Table 5.4. Decision rules in the adaptive test with statistical estimation and testing at  $k_{\max}$ 

Stat. Estimation	Stat. Testing	Decision
if	if	
$\hat{\theta}_{k_{\max}} < -.13$	$\sum_{i=1}^{k_{\max}} a_i x_i \leq \frac{-\sum_{i=1}^{k_{\max}} \ln \left( \frac{q_i(-.13 + \delta)}{q_i(-.13 - \delta)} \right)}{2\delta}$	level 1
$-.13 \leq \hat{\theta}_{k_{\max}} < .33$	$\frac{-\sum_{i=1}^{k_{\max}} \ln \left( \frac{q_i(-.13 + \delta)}{q_i(-.13 - \delta)} \right)}{2\delta} < \sum_{i=1}^{k_{\max}} a_i x_i < \frac{-\sum_{i=1}^{k_{\max}} \ln \left( \frac{q_i(.33 + \delta)}{q_i(.33 - \delta)} \right)}{2\delta}$	level 2
$\hat{\theta}_{k_{\max}} \geq .33$	$\sum_{i=1}^{k_{\max}} a_i x_i \geq \frac{-\sum_{i=1}^{k_{\max}} \ln \left( \frac{q_i(.33 + \delta)}{q_i(.33 - \delta)} \right)}{2\delta}$	level 3

All item selection methods, described before, are involved in the comparisons on  $k$  and  $\%$ . The performance of the item selection methods was also evaluated with respect to the exposure rates of the items and the distribution of the items used over the three subdomains.

The simulations were conducted as follows. An ability of a simulee  $\theta_v$  is randomly drawn from  $N(.294, .522)$ : the ability distribution estimated from the calibration. The three starting items were selected according to the starting procedure discussed earlier and the next items were selected using one of the item selection methods. The simulee's response to an item was generated according to the IRT model. To be more specific: at each exposure, a random number  $g$  was drawn from the interval  $(0,1)$ . For simulee  $v$  and item  $i$ , formula (1) was evaluated and, if:  $p_i(\theta_v) \geq g$ , the item was scored 'correct':  $x_i = 1$ , if not, it was scored 'incorrect':  $x_i = 0$ . This procedure was repeated for  $N=5000$  (first part of the simulation study) and  $N=1000$  (main part of the study) simulees. Besides this general simulation setup, the following variation was also used: 500 test administrations were simulated for 67 equidistant points on the ability scale.

The range on the ability scale was about 2 standard deviations around the mean ability (.294) of the population:  $-0.806 \leq \theta \leq 1.394$ .

## 5.7 Results of the simulation studies

### 5.7.1 Measurement accuracy with statistical estimation

Figure 5.3 portrays the measurement inaccuracy as a function of the test length of the adaptive test using statistical estimation for the various item selection methods. As expected for all selection methods, the inaccuracy decreases as the number of items increases. It is clear that random item selection leads to the greatest inaccuracy. For all the other item selection methods, the decrease in inaccuracy becomes very small after 20 or more items, which justifies the conclusion that the practical requirement of a maximum test length of 25 items is a realistic one.

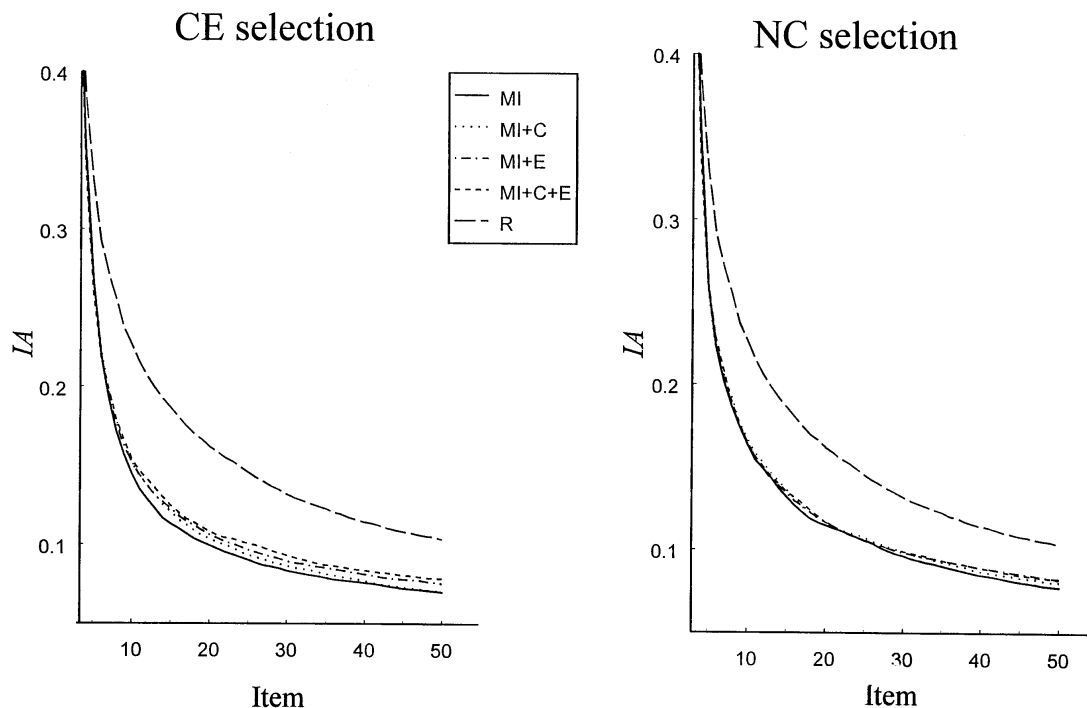


Figure 5.3. Inaccuracy of item selection methods as a function of the number of items

The differences between the item selection methods are small. The inaccuracy with NC selection is a bit larger than with CE selection. For instance, for a test with 20 items with CE selection with content control  $IA = 0.1037$  and with NC selection  $IA = 0.1166$ . Both with CE and NC selection, selecting without constraints (MI) leads to the most accurate estimates. Exposure control has a slightly more negative effect on accuracy than content control. If both constraints are operative, the loss of accuracy is greatest.

### ***5.7.2 The algorithms in the conditions of the placement test***

The results of the simulations with the MI item selection methods are summarized in Table 5.5. The table shows the mean number of items required for taking a decision ( $k$ ) and the percentage of correct decisions (%). To facilitate the interpretation of these results, the administration of the paper-and-pencil version of the placement test was also simulated. With a fixed test length of 25, this resulted in 87.0% correct decisions. Furthermore, it can be noted that with the sample sizes used, differences between the mean number of required items of about 0.6 are statistically significant at the 5% level, whereas differences between percentages of correct decisions are not statistically significant at this level until they are at least 2.4.

#### **Statistical estimation.**

Consider the effect of varying the levels of the preset accuracy in the estimation computation procedure. Increasing the accuracy level results in a noteworthy increase of the mean number of required items, both in CE selection and in NC selection. In CE selection, the increase in the mean number of items used by the various item selection methods varies between 1.9 and 2.8, whereas this effect is about twice as large (between 4.2 and 4.6) in NC selection. In the case of CE selection, the percentages of correct decisions increase considerably (between 2.3 and 4.2%) when the level of accuracy is increased. In the case of NC selection



the gain in the percentages of correct decisions is less (between 1.0 and 2.9%), when the accuracy is increased.

*Table 5.5 Mean number of required items and percentage of correct decisions*

Computation Procedure	Selection Method							
	MI		MI+C		MI+E		MI+C+E	
	k	%	k	%	k	%	k	%
Stat.Estimation								
70%-CE	13.8	85.4	13.8	85.7	14.5	85.5	14.2	83.5
90%-CE	16.3	89.1	16.6	88.8	16.4	87.8	16.4	87.7
70%-NC	14.4	88.4	14.8	88.2	14.3	87.4	14.3	86.3
90%-NC	18.7	89.9	19.4	89.8	18.7	88.4	18.5	89.2
Stat.Testing								
5%- $\delta=.1$	17.6	90.6	18.4	88.9	18.5	86.8	18.7	88.4
10%- $\delta=.1$	15.2	89.5	15.3	88.5	15.9	87.6	16.1	90.0
7.5%- $\delta=.1$	16.7	87.8	16.6	91.1	17.0	88.0	17.8	89.2
7.5%- $\delta=.1333$	14.3	89.1	13.9	89.4	14.6	87.8	14.9	87.4

Only if 90% confidence intervals are used, the percentages of correct decisions are always larger than with the paper-and-pencil version of the placement test. This is in particular true for CE selection. With NC selection, the paper-and-pencil percentage of 87% is also reached in almost all cases.

If CE selection and NC selection are compared, it appears that, if 70% confidence intervals are used, for the NC selection the percentages of correct decisions are higher (between 1.9 and 3.0%), whereas there are small differences in the mean number of required items (the maximum difference is 1 item more with NC selection). With 90% intervals, however, the advantage of NC selection with respect to the number of correct decisions is smaller (between 0.6 and 1.5%), while there is a noteworthy advantage for CE selection (a reduction between 2.3 and 3.1) in the mean number of required items.

Comparing item selection methods, using varying constraints, reveals hardly any differences. Noteworthy differences only occur in comparison to the random item selection method (not included in Table 5.5) which, using confidence intervals of 70% and 90%, led to respectively simulation results:  $k=16.2$  and  $\%=81.7$ ,  $k=20.7$  and  $\%=83.8$ .

### Statistical testing.

If statistical testing is applied as a computation procedure, with a fixed indifference zone of  $\delta = .1$ , the mean number of required items increases considerably when the preset acceptable decision error rates are lowered. The differences at acceptable decision error rates of 5% and 10% vary between 2.4 and 3.1. Unexpectedly, there are hardly any differences between the percentages of correct decisions. This can be explained by the fact that, in a relatively high number of simulated test administrations, a decision could not be taken until 25 items had been administered and thus not on the basis of set error rates; the procedure was stopped by taking the most reasonable decision (see Table 5.4).

If the indifference zone is extended to  $\delta = .1333$ , in the case of acceptable decision error rate of .075, a statistically significant decrease is seen in the mean number of required items (varying between 2.4 and 2.9) without any effect on the percentage of correct decisions. Compared to the paper-and-pencil version of the placement test, the reported simulations in which statistical testing is used as a computation procedure show an increase in the percentage of correct decisions.

Just as in the case of statistical estimation, the differences between the item selection methods are small: only a small increase can be observed in the mean number of required items as the constraints on item selection are tightened. If, in the statistical testing procedure, the items are selected randomly (not included in Table 5.5), this has an effect on the quality of the testing algorithm: a mean number of about 5 additional items is required and the percentage of correct decisions decreases to about 84%.

Comparison of statistical estimation and statistical testing.

The comparison of the testing algorithms using statistical estimation and statistical testing is based on the results in the second, the fourth, and the eighth row of Table 5.5. These three algorithms lead to about equal percentages of correct decisions which all exceed that of the paper-and-pencil test (87.0%). The conclusion that can be drawn from this is that the mean number of required items is smallest in the case of statistical testing as a computation procedure with the following characteristics: the acceptable decisions error rate is  $\alpha = .075$  and the indifference zone is  $\delta = .1333$ . On average, this algorithm requires just over 2 items less than the algorithm that combines statistical estimation as a computation procedure with selection of items with MI at the current ability estimate (CE); another two additional items are required if estimation and selection at the nearest cutting point (NC) is used. If content control and/or exposure control are added as extra constraints to the item selection method, the results are the same.

An attempt was made to find out at what points in the ability distribution the largest reduction in the mean number of required items is to be expected for the three testing algorithms compared. For that purpose, 500 test administrations were simulated at 67 equidistant points on the ability scale,  $-0.806 \leq \theta \leq 1.394$ . Because constraints on the item selection methods had no effect on the outcomes, only the results for the MI selection method are shown. Figure 5.4 gives the percentage of correct decisions as a function of  $\theta$ . The basis for comparison of the three algorithms, about equal percentages of correct decisions, is confirmed by the fact that there are hardly any differences at every value of  $\theta$ .

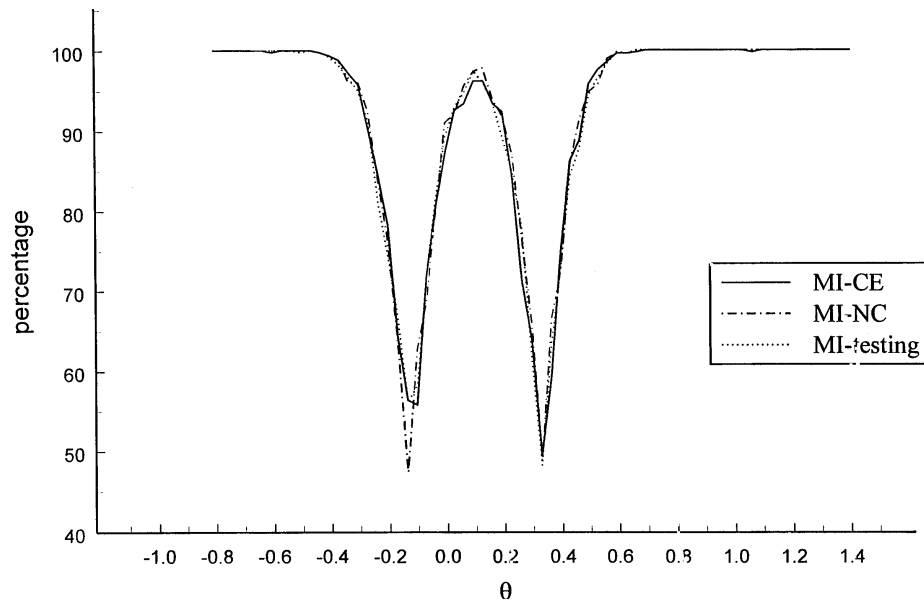


Figure 5.4. Percentage correct decisions for MI item selection as a function of  $\theta$ .

In Figure 5.5, the mean number of required items is depicted as a function of  $\theta$ . For all three algorithms, the abilities between the cutting points (-.13 and .33) required generally the largest number of items. Very able students required more items than students with lower abilities. This may be a consequence of the content of the item bank, which contains a relatively high number of easy items (see Figure 5.1). A more obvious cause, however, could also be the starting procedure used in the adaptive test, which begins with three easy items.

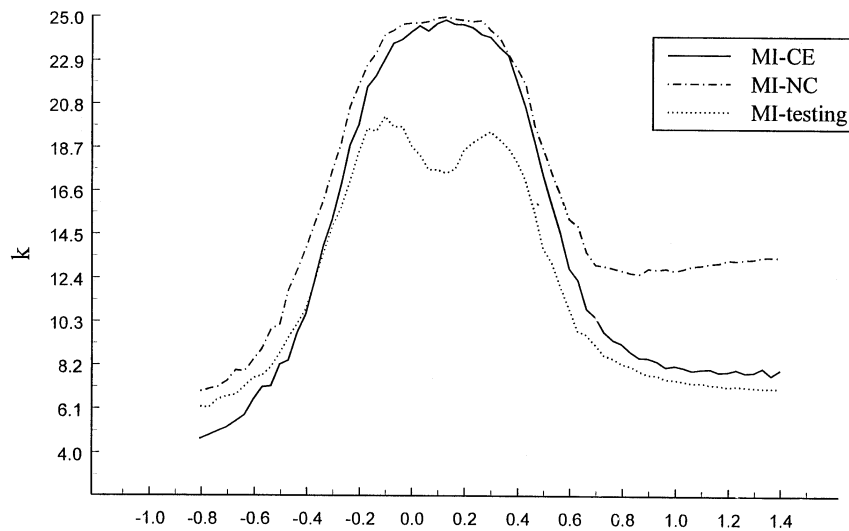


Figure 5.5. Mean number of required items for MI item selection as a function of  $\theta$ .

A comparison of the three algorithms shows that selecting items with MI at the current ability estimate (CE) requires less items than selection at the nearest cutting point (NC) for all abilities. In particular this is true for students with higher abilities. Comparing the more efficient estimation procedure (CE) to the testing procedure shows a reduction in the mean number of items required for classification for the testing procedure, especially for the abilities between the cutting points. It is only for the lower abilities that the estimation procedure with MI-CE performs slightly better.

Exposure Data. It was previously concluded that imposing constraints on the item selection method has no serious consequences for the quality of the testing algorithms in the conditions of the placement test. The question now is: Do the constraints imposed have the desired effects?

Table 5.6 shows the exposure rates of the items from the item bank for various item selection methods. Reported is the number of items (from a total number of 250) with the frequency of use in percentages ( $f$ ) in 1000 simulated test administrations. The results are given for the testing algorithms using statistical

estimation with 90% confidence intervals, and for the algorithms using statistical testing with  $\alpha=.075$  and an indifference zone of  $\delta=.1333$ .

Table 5.6. Exposure rates of items with statistical estimation ( $\gamma=1.644$ ) and with statistical testing ( $\alpha=.075$ ;  $\delta=.13$ ) as computation procedure

freq. %	Selection Method												
	R	MI			MI+C			MI+E			MI+C+E		
		CE	NC	T	CE	NC	T	CE	NC	T	CE	NC	T
$f=0$	0	132	156	156	126	156	156	75	103	104	53	75	77
$0 < f \leq 2.5$	0	5	1	1	11	2	2	47	21	27	56	40	43
$2.5 < f \leq 5$	0	28	28	28	31	29	28	26	27	32	42	36	38
$5 < f \leq 7.5$	130	26	14	17	16	14	16	20	21	18	17	20	25
$7.5 < f \leq 10$	68	8	4	10	18	4	13	13	11	13	18	12	11
$10 < f \leq 15$	43	17	11	12	11	6	12	26	11	27	19	18	23
$15 < f \leq 20$	9	13	3	7	5	3	6	24	26	11	27	22	16
$20 < f \leq 25$	0	2	8	4	13	7	1	8	11	6	9	10	7
$25 < f \leq 30$	0	2	5	2	4	6	4	7	4	2	5	4	2
$30 < f \leq 40$	0	7	7	2	7	10	4	4	10	7	4	7	5
$40 < f \leq 50$	0	4	1	4	4	2	2	0	5	3	0	5	3
$50 < f \leq 75$	0	6	9	5	3	5	4	0	0	0	0	1	0
$75 < f \leq 100$	0	0	3	2	1	6	2	0	0	0	0	0	0

The item bank is most efficiently used by the random item selection method: all items are used in between 5% and 20% of the test administrations. From a measurement point of view, the better selection methods suffer from underutilization as well as from overutilization of parts of the item bank. With statistical estimation, MI-CE selection, for instance, 132 items are never used at all, whereas 21 items are used frequently (in over 20% of the administrations). The latter could become problematic with regard to the confidentiality of the items. If statistical estimation is used, the comparison of the NC selection methods with the CE selection methods shows that all variants of NC selection

make less efficient use of the item bank: the number of items never used as well as the number of items used frequently is larger.

Comparing the exposure rates in the case of statistical testing as a computation procedure with the case of statistical estimation combined with the NC selection method shows that the first method uses the item bank better. There are no differences in the number of never used items, but the number of frequently used items is about 1.5 times larger in the NC procedure (e.g., 17 against 27 items with option MI+C+E). However, compared to statistical estimation with CE selection, the exposure rates with statistical testing as a computation procedure are less favorable.

For the effects of content and exposure control on the exposure rates, we concentrate on the CE item selection method (columns 3,6,9,12 in Table 5.6). Applying content control has a notable effect: the number of items that is never used decreases slightly (from 132 to 126), but there is an increase in the number of items used frequently (from 21 to 32). Applying exposure control has the expected positive effect on the number of items not used (75); the number of items frequently used also decreases to 19. Moreover, there are no items that are used in more than 40% of the test administrations. Combining content and exposure control clearly has the most positive effect on the exposure rates of the items.

Finally, Table 5.7 shows the distributions of the items used over the three content subdomains for the selection methods reported in Table 5.6. It appears that the desired distribution - from the point of view of content balancing - over subdomains A, B, and C (16:20:64) can be achieved only through explicit content control in selecting items. All other selection methods over-represent subdomain A and under-represent subdomain C by about 4% in the average test.

Table 5.7 Distribution of items used over subdomains

Selection Method	subdomains (desired %)		
	A (16%)	B (20%)	C (64%)
Statistical Estimation			
R	21.2	21.7	57.2
CE-MI	21.3	20.0	58.7
NC-MI	20.9	15.6	63.6
CE-MI+C	16.2	20.8	62.9
NC-MI+C	16.2	20.4	63.4
CE-MI+E	22.6	21.0	56.4
NC-MI+E	22.7	18.7	58.7
CE-MI+C+E	16.3	20.9	62.8
NC-MI+C+E	16.5	20.4	63.1
Statistical Testing			
R	21.2	21.2	57.1
MI	23.5	14.5	62.0
MI+C	16.7	21.5	61.8
MI+E	23.2	18.6	58.1
MI+C+E	16.6	21.0	62.3

## 5.8 Discussion

The performance of the statistical computation procedures and the item selection methods for the three-way classification problem were studied by means of simulation studies on an operational item bank. These studies lead to the following conclusions with regard to the development of the computerized adaptive placement test for mathematics:

1. The quality of the item bank is satisfactory for the purpose of adaptive testing;
2. The absolute maximum of 25 items for each test administration is realistic;



3. The reduction in the number of required items can be expected to amount to between 22% and 44% of the number of items in the paper-and-pencil version of the placement test;
4. Applying the double SPRT is the most promising computation procedure in the testing algorithm;
5. Additional constraints on item selection methods in the form of content control or a mild form of exposure control can be imposed without impairing the quality of the procedures; and
6. Before deciding on a final implementation of a CAT in the placement test for mathematics, it should be determined experimentally whether the way the algorithms operate in real testing situations is in line with the results of the simulations.

In general, the following conclusions are drawn. With regard to the testing algorithms used for the classification of examinees into three categories, the conclusion is that statistical testing as a computation procedure is a promising alternative to the more traditional statistical estimation procedure. Apart from the gain in the mean number of required items, while attaining approximately equal accuracy, statistical testing has the added advantage of little computational work during the test administration. In statistical estimation, an iterative maximalization procedure has to be followed; in statistical testing, in the IRT model used, a simple comparison of the observed weighted score with constants suffices. As in the two way classification problem, in the three way classification problem statistical testing in a CAT algorithm also offers a relatively simple and sound procedure which is performing better than traditional tests. Results reported in this study on the relation between the size of the acceptable decision error rates and the width of the indifference zone and the performance of the test are consistent with those of Reckase (1983) and Spray and Reckase (1996) for the two-way classification. It must be noted, however, that, in the present study the comparison between statistical testing and estimating does not have a proper statistical basis, and therefore the generalizability of our results is not guaranteed.

The problem is that in estimation, the classification accuracy is determined by  $\gamma$ , and in testing, by  $\alpha$  and  $\delta$ . As there are no formal relationships between these parameters, which could ensure that both procedures lead to the same classification accuracy, the comparisons made between mean number of required items give only indications. The approach followed in this study resembles the one followed by Kingsbury and Weiss (1983) for the classification into two categories. Spray and Reckase (1996) recently presented a better way to compare statistical testing and estimating for the two-way classification: in their study they proposed a procedure that leads to a matching of the accuracy of testing and estimating. It would be worthwhile to investigate whether their procedure is generalizable to the classification into three categories. This general problem, as well as the consequences of truncating all the algorithms at a maximum test length for the acceptable decision error rates in relation to the width of the indifference zones, and the quality of the testing algorithms, all call for further research.

It is interesting to note that, in contrast to the findings of Spray and Reckase (1994) for the two-way classification problem, in the current study, when statistical estimation was used as a computation procedure, it was advisable to select items that have maximum information at the current ability estimate rather than items that are maximally informative at the nearest cutting point. Whether this is partly due to the characteristics of the item bank used, the cutting points chosen, or the chosen practical requirements, is still a question to be answered. With regard to the item selection methods used in combination with statistical testing as a computation procedure, it can be stated that these can probably be improved. This study has deliberately not used estimates of the examinees' ability in statistical testing as a computation procedure. It is expected that in a follow-up study in which we do resort to estimates and use alternative item selection methods, it will appear that the testing procedure can still be improved.

## 5.9 References

- Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society, B*, 12, 137-144.
- Kingsbury, G.G., & Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D.J. Weiss (Ed.), *New horizons in testing* (pp. 257-286). New York: Academic Press.
- Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*, 2, 359-375.
- Kingsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4, 241-261.
- Lewis, C., & Sheenan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, 14, 376-386.
- Lord, F.M. (1971). A theoretical study of two-stage testing. *Psychometrika*, 36, 227-242.
- Reckase, M.D. (1983). A procedure for decision making using tailored testing. In: D.J. Weiss (Ed.), *New horizons in testing* (pp. 237-255). New York: Academic Press.
- Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics*, 20, 502-522.
- Spray, J.A., & Reckase, M.D. (1994, April). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.
- Spray, J.A., & Reckase, M.D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, 21, 405-414.

- Sympson, J.B., & Hetter, R.D. (1985). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the annual conference of the Military Testing Association, San Diego.
- Thissen, D, & Mislevy, R.J. (2000). In Wainer, H., Dorans, N.J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L.,& Thissen, D. (Eds.), *Computerized adaptive testing: A primer. Second Edition* (pp.101-133). Hillsdale, NJ: Lawrence Erlbaum.
- Verhelst, N.D., & Glas, C.A.W. (1995). The one-parameter logistic model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp.215-237). New York: Springer-Verlag.
- Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *One-parameter logistic model (OPLM)*. Arnhem: Cito.
- Wainer, H (2000). Introduction and history. In Wainer, H., Dorans, N.J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L.,& Thissen, D. *Computerized adaptive testing: A primer. Second Edition* (pp.1-21). Hillsdale, NJ: Lawrence Erlbaum.
- Wald, A. (1947). *Sequential analysis*. New York: Wiley.
- Warm, T.A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.
- Weiss, D.J., & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361-375.



## Chapter 6

### Item selection in adaptive testing with the sequential probability ratio test<sup>1</sup>

---

<sup>1</sup>This chapter is a minor revised reprint of: Eggen, T.J.H.M.(1999). Item selection with adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, 23, 249-261.

## **Abstract**

Computerized adaptive tests (CATs) were originally developed to obtain an efficient estimate of an examinee's ability. For classification problems, applications of the Sequential Probability Ratio Test (Wald, 1947) have been shown to be a promising alternative for testing algorithms which are based on statistical estimation. However, the method of item selection currently being used in these algorithms, which use statistical testing to infer on the examinees, is either random or based on a criterion which is related to optimizing estimates of examinees (maximum (Fisher) information).

In this study, an item selection method based on Kullback-Leibler information is presented, which is theoretically more suitable for statistical testing problems and which can improve the testing algorithm for classification problems.

Simulation studies were conducted for two- and three-way classification problems, in which item selection based on Fisher information and Kullback-Leibler information were compared. The results of these studies show that the performance of the testing algorithms with Kullback-Leibler information-based item selection are sometimes better and never worse than algorithms with Fisher information-based item selection.

## **6.1 Introduction**

Efficient estimation of the ability of an examinee is often the purpose of computerized adaptive testing (CAT). But, if the goal of testing is to classify examinees in a limited number of categories, for example, pass/fail decisions on an exam or decisions regarding placement in courses on different levels, CATs can make use of algorithms that are based on statistical testing rather than statistical estimation. Studies by Reckase (1983), Lewis and Sheenan (1990), and Spray and Reckase (1994, 1996), for decisions in two categories, and Eggen and Straetmans (2000), for decisions in three categories, have shown that the Sequential Probability Ratio Test (SPRT) (Wald, 1947) can be successfully applied in adaptive testing using an item response theory (IRT) calibrated item bank. See also chapter 5 of this thesis.

An important part of a CAT algorithm is the item selection procedure, which determines, during testing, the choice of the items which are administered. In current adaptive tests using statistical testing in the algorithm, item selection is based on a criterion which is closely related to statistical estimation. Items are selected that maximize the item Fisher information, which means the item will be chosen that minimizes the expected contribution of an item to the standard error of the ability estimate of an examinee. In this chapter, item selection procedures will be proposed that are based on Kullback-Leibler information (Cover & Thomas, 1991). It will be shown that the item Kullback-Leibler information expresses the expected contribution of an item to the discriminatory power between two hypotheses. Conceptually, K-L information fits the statistical testing algorithm more closely than Fisher information. One of the questions addressed in this study is whether using K-L information has a positive impact on the performance of the adaptive tests with statistical testing for decision problems with two categories and with three categories. Bayesian item selection criteria are also in use in adaptive testing with estimation. These criteria, recently discussed by Van der Linden (1998), will not be considered in this study.

The first part of the chapter is an overview of the SPRT application in problems with two and three categories. Next, item selection based on both Fisher and



Kullback-Leibler information will be presented. Finally, a comparison of the item selection procedures for both the two- and three- category problem will be made on the basis of simulation studies with an operational item bank.

## 6.2 Sequential testing in the testing algorithm

In testing algorithms of adaptive tests, the likelihood function of an examinee's ability,  $\theta$ , plays a central role in the inference on the examinee. Assuming that an IRT calibrated item bank is available, that is, the parameters of the items can be considered to be known, and given the scores on  $k$  items ( $x_i, i=1,...,k$ ), this function is

$$L_k(\theta; x_1, \dots, x_k) = L_k(\theta; \underline{x}) = \prod_{i=1}^k L(\theta; x_i) = \prod_{i=1}^k p_i(\theta)^{x_i} (1 - p_i(\theta))^{1-x_i}. \quad (1)$$

In this likelihood function,  $p_i(\theta)$ , the probability of answering item  $i$  correctly, is the item response function belonging to an IRT model. In this study, the two-parameter logistic (2-PL) model is used:

$$p_i(\theta) = P(X_i = 1 \mid \theta) = \frac{\exp a_i(\theta - b_i)}{1 + \exp a_i(\theta - b_i)}. \quad (2)$$

The response to an item  $x_i$  is either correct (1) or incorrect (0). The probability of a correct response increases with the latent ability  $\theta$  and depends on two item characteristics: the difficulty parameter,  $b_i$ , and the discrimination parameter,  $a_i$ .

In traditional adaptive tests, the ability is estimated after each item by maximizing the likelihood function with respect to  $\theta$ . When statistical testing rather than estimation is used in the testing algorithm, the likelihood function is used somewhat differently, which will become clear in the following description of the statistical testing procedure.

### 6.2.1 Classification in two categories

On the latent ability scale, a decision or cutting point  $\theta_0$  between, for example, a master and non-master, or between an examinee who passes and an examinee who fails on an exam, is given. A small region on both sides of this point, a so-called indifference zone, is selected. The width of these regions, although they could differ from each other, is taken to be  $\delta$ . The indifference interval expresses the fact that, owing to measurement errors, making the right decision about examinees very near the cutting point can never be guaranteed. One could also say that the interval expresses the indifference of an examiner of the classification of the examinees who are that near to the cutting point.

Next, the statistical hypotheses are formulated:

$$H_0: \theta \leq \theta_0 - \delta = \theta_1 \text{ against } H_1: \theta \geq \theta_0 + \delta = \theta_2. \quad (3)$$

Acceptable decision error rates are specified as follows:

$$P(\text{accept } H_0 \mid H_0 \text{ is true}) \geq 1 - \alpha, \text{ and } P(\text{accept } H_0 \mid H_1 \text{ is true}) \leq \beta, \quad (4)$$

in which  $\alpha$  and  $\beta$  are small constants. The test meeting these decision error rates can be carried out using the SPRT (Wald, 1947). The test statistic used is the ratio between the values of the likelihood function (1) under the alternative hypothesis and the null hypothesis:

$$LR_k(\theta_2, \theta_1; \underline{x}) = \frac{L_k(\theta_2; \underline{x})}{L_k(\theta_1; \underline{x})}. \quad (5)$$

It will be clear that high values of this ratio indicate the examinee is more likely to have an ability above the cutting point, and small values support the decision that the examinee is below the cutting point. That is, the test meets the error rates if the following procedure is used:

$$\text{Continue sampling if: } \beta/(1 - \alpha) < LR_k(\theta_2, \theta_1; \underline{x}) < (1 - \beta)/\alpha; \quad (6)$$

$$\text{accept } H_0 \text{ if: } LR_k(\theta_2, \theta_1; \underline{x}) \leq \beta/(1 - \alpha); \quad (7)$$

$$\text{reject } H_0 \text{ if: } LR_k(\theta_2, \theta_1; \underline{x}) \geq (1 - \beta)/\alpha. \quad (8)$$

Equation (6) is called the critical inequality of the test. It can easily be shown that if the 2-PL model (2) is used, the critical inequality of this test can be written as

follows:

$$\frac{\ln \beta / (1 - \alpha) - C_{k\theta_1\theta_2}}{\theta_2 - \theta_1} < \sum_{i=1}^k a_i x_i < \frac{\ln (1 - \beta) / \alpha - C_{k\theta_1\theta_2}}{\theta_2 - \theta_1}. \quad (9)$$

$$\text{In (9), } C_{k\theta_1\theta_2} = \sum_{i=1}^k \ln \frac{1 + \exp a_i(\theta_1 - b_i)}{1 + \exp a_i(\theta_2 - b_i)} = \sum_{i=1}^k \ln \frac{1 - p_i(\theta_2)}{1 - p_i(\theta_1)} = \sum_{i=1}^k \ln \frac{q_i(\theta_2)}{q_i(\theta_1)}, \quad (10)$$

which only depends on the item parameters and on constants in the statistical testing procedure,  $\theta_1$  and  $\theta_2$ , that are chosen beforehand. The evaluation of the critical inequality is quite easy because it involves only the observed weighted score (in case of the 2PL model) and known constants. Note that because  $\theta_2 > \theta_1$ ,  $q_i(\theta_2)/q_i(\theta_1) < 1$  and thus  $C_{k\theta_1\theta_2} < 0$ . Furthermore, if the indifference interval  $2\delta = \theta_2 - \theta_1$  increases, the width of the critical interval gets smaller, which indicates that shorter tests can be used to make a decision.

Although Wald (1947) has shown that eventually a decision will be taken with probability 1 with the SPRT, in practice, a maximum test length,  $k_{\max}$ , is usually specified. At this test length, a forced decision is taken. In that case, the most obvious decision is taken:  $H_0$  is rejected if the test statistic is larger than the midpoint of the critical inequality interval; otherwise it is accepted.

### 6.2.2 Classification in three categories

The above testing procedure is readily generalized to cases of classification in one of three categories. In this case, there are two cutting points,  $\theta_1$  and  $\theta_2$ , and three different levels of ability are distinguished. See Figure 6.1.

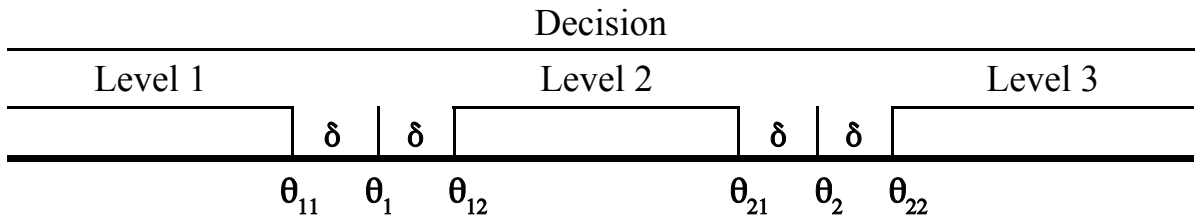


Figure 6.1 Schematic representation of the classification problem with three categories.

After selecting the indifference zones, all taken to be  $\delta$ , two pairs of hypotheses are formulated:

$$H0\_1: \theta \leq \theta_{11} = \theta_1 - \delta \text{ (level 1)} \quad H1\_1: \theta \geq \theta_{12} = \theta_1 + \delta \text{ (higher than 1)} \quad (11)$$

$$H0\_2: \theta \leq \theta_{21} = \theta_2 - \delta \text{ (lower than 3)}; \quad H1\_2: \theta \geq \theta_{22} = \theta_2 + \delta \text{ (level 3)}. \quad (12)$$

The SPRT test described in the preceding section is applied for each pair of hypotheses. In the specification of the acceptable decision errors, as in Equation 4, the small constants  $\alpha_1$  and  $\beta_1$ ,  $\alpha_2$  and  $\beta_2$ , respectively are used.

If the 2-PL model is used, the critical inequalities of the tests are

$$L_1 = \frac{\ln \frac{\beta_1}{1 - \alpha_1} - \sum_{i=1}^k \ln \frac{q_i(\theta_1 + \delta)}{q_i(\theta_1 - \delta)}}{2\delta} < \sum_{i=1}^k a_i x_i < \frac{\ln \frac{1 - \beta_1}{\alpha_1} - \sum_{i=1}^k \ln \frac{q_i(\theta_1 + \delta)}{q_i(\theta_1 - \delta)}}{2\delta} = U_1 \quad (13)$$

$$L_2 = \frac{\ln \frac{\beta_2}{1 - \alpha_2} - \sum_{i=1}^k \ln \frac{q_i(\theta_2 + \delta)}{q_i(\theta_2 - \delta)}}{2\delta} < \sum_{i=1}^k a_i x_i < \frac{\ln \frac{1 - \beta_2}{\alpha_2} - \sum_{i=1}^k \ln \frac{q_i(\theta_2 + \delta)}{q_i(\theta_2 - \delta)}}{2\delta} = U_2 \quad (14)$$

It can easily be checked that  $\partial (\ln q_i(\theta + \delta)) / (q_i(\theta - \delta)) / \partial \theta < 0$  for all  $\theta$ , which means it is monotone decreasing in  $\theta$ . A consequence of this is that the lower bound of test 1,  $L_1$ , can never be larger than the upper bound of test 2,  $U_2$ . So, by combining the decisions of the simultaneously applied two SPRTs, unequivocal decisions can be made in the three-way classification problem.

The decisions based on a combination of two SPRTs are:

decision test 2	decision test 1	
	1	2 or 3
1 or 2	1	2
3	impossible	3

This generalization of the SPRT, known as Sobel and Wald's (1949) combination procedure, performs as well as Armitage's (1950) combination procedure, which is applied by Spray (1993) for classification in three and even  $k$  categories.

The procedure for the combined test with the 2-PL model is as follows:

$$\text{take decision 1 if } \sum_{i=1}^k a_i x_i \leq L_1; \quad (15)$$

$$\text{take decision 2 if } U_1 \leq \sum_{i=1}^k a_i x_i \leq L_2; \quad (16)$$

$$\text{take decision 3 if } \sum_{i=1}^k a_i x_i \geq U_2; \quad (17)$$

$$\text{continue testing if } \text{else.} \quad (18)$$

A sketch of the procedure is given in Figure 6.2.

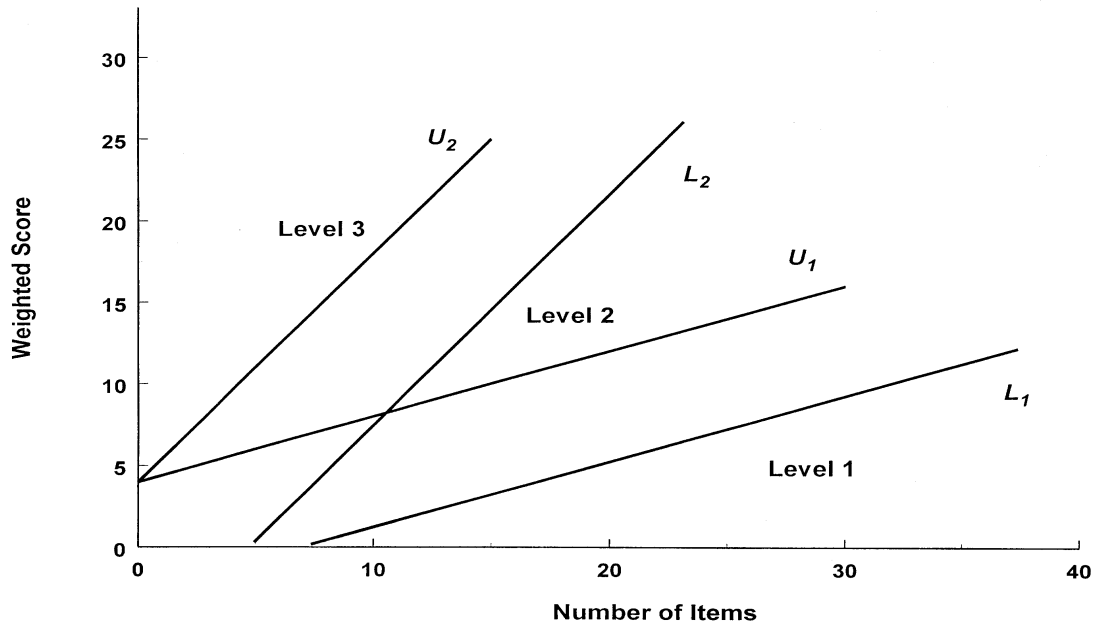


Figure 6.2. Sketch of the statistical test procedure with three levels.

Note that, of course,  $L_1 < U_1$  and  $L_2 < U_2$ , but it generally requires some items before  $U_1 < L_2$  can be true and decision 2 can be taken.

### 6.3 Item selection

In the testing algorithm, the item selection procedure chooses items from the item bank. In connection with the use of the SPRT, random selection is a possibility, but it is well known that the efficiency is much greater when a maximum information selection strategy is applied (see, for example, Eggen & Straetmans, 2000). Information usually means Fisher information which will be described first.

In adaptive testing in which the aim is estimating the ability of an examinee, items are selected to have maximum Fisher information at the current ability estimate. Chang and Ying (1996) introduced an item information measure which is not based on Fisher information but on Kullback-Leibler information (K-L information). In this section, an information measure will be introduced that is also based on Kullback-Leibler information but which is more suitable to be used in connection with adaptive testing with statistical testing with the SPRT.

Some item selection procedures for both Fisher information and Kullback-Leibler information will be given for both the two- and three-categories problem.

#### 6.3.1 Fisher information

Current maximum information item selection procedures are almost all based on the Fisher item information, which for an item  $i$  is defined as

$$I_i(\theta) = \mathcal{E} \left( \frac{\frac{\partial}{\partial \theta} L(\theta; x_i)}{L(\theta; x_i)} \right)^2. \quad (19)$$

In the 2-PL, the expression is given by

$$I_i(\theta) = a_i^2 p_i(\theta) q_i(\theta). \quad (20)$$

For a test consisting of  $k$  items, the test information is the sum of the information of the items in the test:  $I(\theta) = \sum_{i=1}^k I_i(\theta)$ . Selecting items with maximum information maximizes the contribution to the test information. The

usefulness of this is readily understood if an estimate of the ability of an examinee is wanted, especially when the maximum likelihood estimator (MLE) is used. In this case,  $\hat{\theta}_k$ , the MLE after taking  $k$  items, follows from  $\max_{\theta} \prod_{i=1}^k L(\theta; x_i)$  and its standard error is estimated by  $se(\hat{\theta}_k) = 1 / \sqrt{I(\hat{\theta}_k)}$ . So, it can be seen that by selecting items having maximum information, the contribution to the decrease of the standard error is greatest. Furthermore, from the definition in (19), it can be seen that maximizing the information is the same as maximizing the contribution of an item to the expected relative rate of change of the likelihood function. As Chang and Ying (1996) point out, the greater this change rate at a given value of  $\theta$ , the better it can be distinguished from points near to this value, and the better this value can be estimated.

#### Some selection procedures based on Fisher information.

- F1 In adaptive tests in which the examinee's ability is to be estimated, the most popular item selection method is to select the item that has maximum information at the current ability estimate:  
Select the item  $i$  for which:  $\max_i I_i(\hat{\theta}_k)$ .
- F2 Spray and Reckase (1994) have shown that in a classification problem with two categories for which the SPRT procedure is being used, it is more efficient to select the items which have maximum information at the cutting point  $\theta_0$  rather than at the current ability estimate:  
Select the item  $i$  for which:  $\max_i I_i(\theta_0)$ .
- F3 In a three-way classification problem for which the generalized SPRT described in the preceding section is being used, an alternative selection method could be to select the item which maximizes the information at the cutting point nearest to the current estimate:  
Determine  $\min(|\theta_1 - \hat{\theta}_k|, |\theta_2 - \hat{\theta}_k|)$  and choose the item with maximum information at the cutting point with the minimum.
- F4 For the three-way classification problem with the SPRT in which no use is made of estimates of abilities, Eggen and Straetmans (2000) propose selecting the item which maximizes the information at the cutting point

## Chapter 7

### Optimal testing with easy or difficult items in computerized adaptive testing<sup>1</sup>

---

<sup>1</sup>The work in this chapter was done in cooperation with A. A. Verschoor. The paper will be published as Measurement and Research Department Report 2004-2 Arnhem: Cito and will be submitted for publication in Applied Psychological Measurement.



## **Abstract**

Computerized adaptive tests (CATs) are individualized tests which, from a measurement point of view, are optimal for each individual, possibly under some practical conditions. In the present study it is shown that maximum information item selection in CATs using an item bank which is calibrated with the one- or the two-parameter logistic model, results in each individual answering about 50% of the items correctly. Two item selection procedures giving easier (or more difficult) tests for students are presented and evaluated. Item selection on probability points of items yields good results only with the 1pl model and not with the 2pl model. An alternative selection procedure, based on maximum information at a shifted ability level, gives satisfactory results with both models

## **7.1 Introduction**

Computerized adaptive tests (CATs) are individualized tests that are administered in an automated environment. CATs are used for estimating the ability of a student or for making a decision on, for instance, the most appropriate training program for that student. It has been shown that, compared to traditional linear tests, CATs yield a considerable gain in efficiency. In the literature (see, e.g., Wainer, 2000 and Eggen & Straetmans, 2000), it has been reported that halving the average number of items needed is feasible, while at the same time the accuracy of the ability estimates or the decisions taken is maintained. CATs make use of item banks which are calibrated using item response theory (IRT) (Hambleton & Swaminathan, 1985). The gain in CATs is realised by selecting, on the basis of the results on previously administered items, the most informative item from an available item bank. During testing, the optimal item is chosen after every item for every student and thus the optimal test is assembled and administered.

CAT-tailored testing has a number of frequently mentioned advantages: the gain in measurement efficiency goes hand in hand with the fact that each student is challenged at his or her own level because items which are too difficult or too easy for a given student will never be administered. Initially, the intended optimality, and, consequently, the item selection method, was based solely on a measurement theoretic or psychometric criterion. The criterion of maximum item information at the current ability estimate is in common use (Van der Linden & Pashley, 2000). The increasing number of CAT applications has resulted in more consideration being given to content-based and practical requirements or conditions in item selection algorithms. Applying content control (Kingsbury & Zara, 1991) and exposure control (Eggen, 2001) is routinely possible. In modern CATs, items that are psychometrically optimal are selected from an item bank which, to the degree possible, meets these practical conditions.

The aim of the present study was to determine whether it is possible to take consideration of testees even more by not only considering the practical conditions, but also by relaxing the psychometrically optimal selection.

Psychometrically optimal selection of items means that items will always be chosen for an individual student, which he or she, at his or her thus far known ability level, has a 50% probability of answering correctly. Thus, as a rule, students taking a CAT will answer about half of the items correctly. Although the difficulty of the items is taken into account in the scoring of a student, it can be the case that CAT tests are perceived as very difficult for each individual student and this could have possible negative side effects, for example, enhanced test anxiety and, consequently, possible lower test performance. This could especially be the case for tests which are administered in primary and secondary education, where, traditionally, tests are constructed in such a way that the average student has, on average, a somewhat higher probability (60 or 70%) of correctly answering the items.

One approach for reducing possible negative effects on the difficulty of the items is self-adaptive testing (Rocklin & O' Donnell, 1987). Self-adapted tests (SATs) are CATs in which the difficulty level of each item is chosen by the examinee rather than by the CAT algorithm. SATs have been studied rather extensively in recent years. The meta-analysis by Pitkin & Vispoel (2001), comparing SATs with CATs, gives an overview. In general, test anxiety reduction is reported in SATs and there is also a little gain in the average performance of examinees if SATs are compared to CATs. This gain could be caused by the reduction of anxiety. Another explanation is that part of the gain could be explained by the, on average, larger bias of the (maximum likelihood) ability estimates in the SATs, as a consequence of self-selection of the items, compared to the CATs. Compared to CATs, SATs are less efficient: more items are needed to reach the same measurement precision. Finally, it can be mentioned that, for students, a SAT is more time consuming than a CAT and that implementing a SAT still entails a number of unresolved problems related to, for example, the exact information the students should be asked and the design of the interface.

In the present chapter, the possibilities for using CATs with selection methods in their algorithms which lead to higher (or lower) success probabilities than 50%

were explored. Changing the CAT algorithm for that reason was also proposed in a study by Bergstrom, Lunz & Gershon (1992). They successfully applied an algorithm which chooses easier items, but only for the case of the one-parameter logistic IRT model. In the present study, for both the one- and the two-parameter logistic IRT model, two CAT selection methods, which choose items with varying difficulties were developed and the consequences for the measurement efficiency evaluated.

## 7.2 Item selection in CAT

Computerized adaptive tests presuppose the availability of an IRT-calibrated item bank. The algorithms for adaptive tests operate on the basis of the item parameters from an IRT model. The IRT model used in this study is the two-parameter logistic model (2pl). In this model, the probability of correctly answering item  $i$ , also called the item response function, is given by

$$p_i(\theta) = P(X_i = 1 \mid \theta) = \frac{\exp(\alpha_i(\theta - \beta_i))}{1 + \exp(\alpha_i(\theta - \beta_i))}$$

where,  $\beta_i$  is the location parameter of the item. This parameter is associated with the difficulty of the item. It is the point on the ability scale at which the student has a 50% chance of correctly answering the item. Parameter  $\alpha_i$  is the item's discrimination parameter. If the discrimination parameter for all items is the same,  $\alpha_i = \alpha$ , this is the special case of the one-parameter logistic model (1pl). In a calibrated item bank, estimates of the values of ( $\alpha_i$  and)  $\beta_i$  for each item have been stored in the bank. After the administration of an item, the next item selected from the item bank is the one that best matches the ability demonstrated by the candidate up to that point. Usually the (Fisher item) information is used for selecting. In the case of the two-parameter model, this function is given by

$$I_i(\theta) = \alpha_i^2 p_i(\theta) (1 - p_i(\theta)) = \frac{\alpha_i^2 \exp(\alpha_i(\theta - \beta_i))}{(1 + \exp(\alpha_i(\theta - \beta_i)))^2}.$$

This item information function expresses the contribution an item can make to the accuracy of the measurement of a person as a function of his or her ability. This becomes clear if one realizes that the estimation error of the ability estimate can be expressed as a function of the sum of the item information of the items administered:

$$se(\hat{\theta}_k) = 1 / \sqrt{\sum_{i=1}^k I_i(\hat{\theta}_k)}$$

Items are selected according to the following procedure: after the ability estimate  $\hat{\theta}_k$  has been determined, the information for each item that has not yet been administered is computed at this point; the item whose information value is highest is then selected and administered.

### The item information function

For dichotomous items, the Fisher item information is a single-peaked function of the ability. In the two-parameter model, it shows that, for each

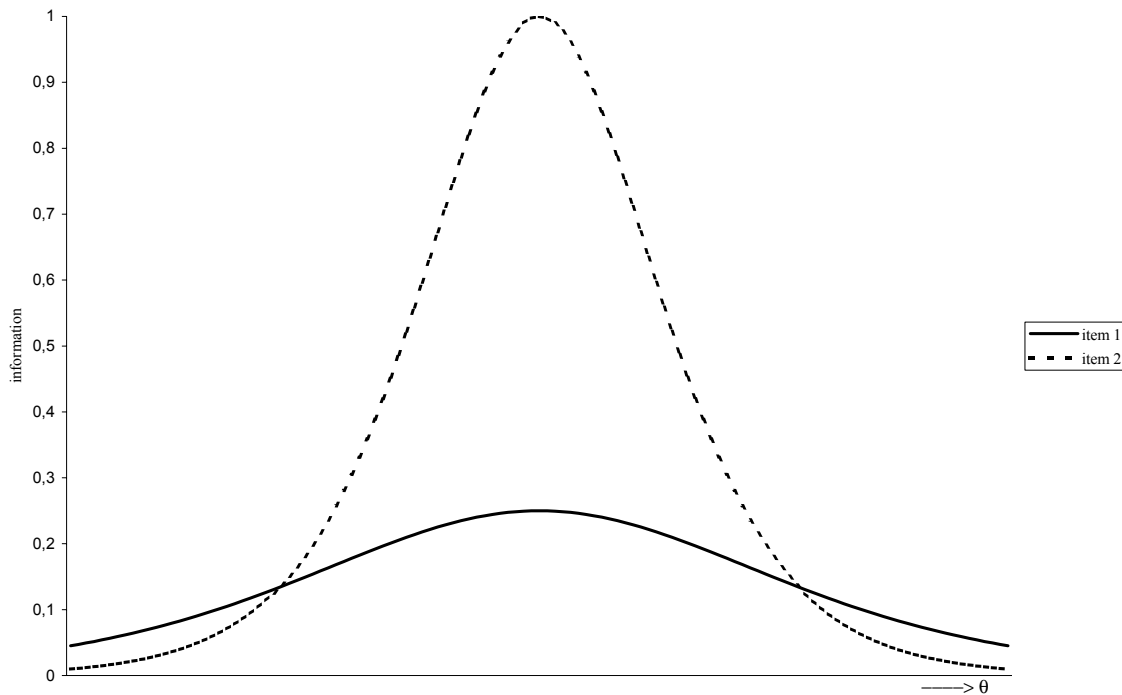


Figure 7.1. Item information functions:  $\beta_1 = \beta_2 = 0$  and  $\alpha_1 = 1$ ,  $\alpha_2 = 2$

item, the information reaches its maximum at the value of the location parameter (difficulty) of the item (  $\theta = \beta_i$  ). In addition, it is clear that the discrimination parameter has a great influence on the information. The larger the  $\alpha_i$ , the greater the information.

The relation between the information in an item and the probability of succeeding on an item for any item following the 1pl or the 2pl model is given in Figure 7.2.

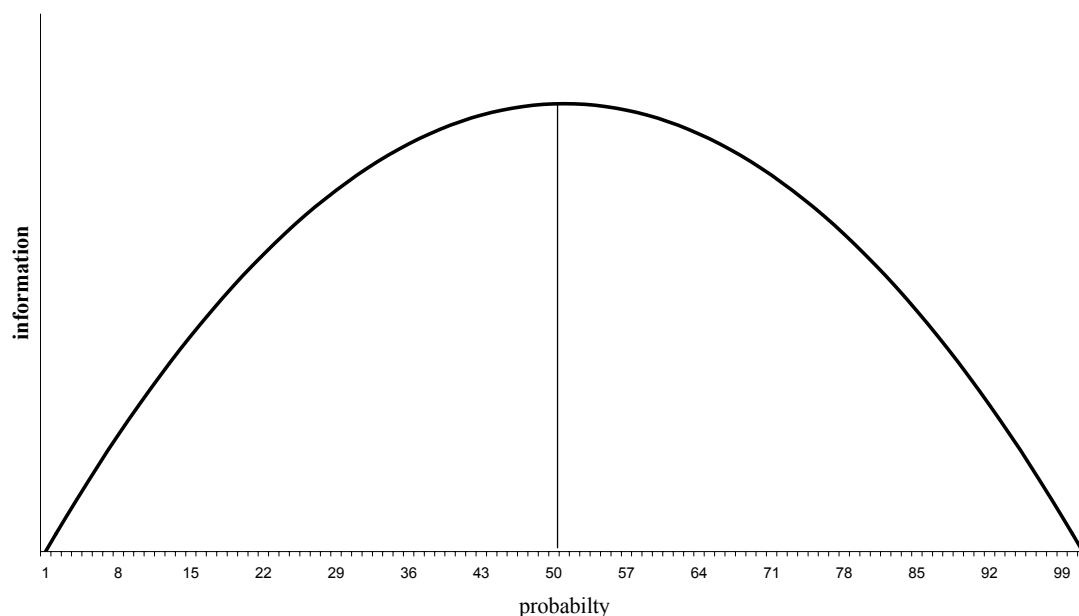


Figure 7.2. Item information as a function of success probability

One can see that an item gives maximum information at a success probability of 0.50. At other probability levels, there is always less information.

### 7.3 Item selection on the basis of success probability

For each item, ability levels can be defined at which there is a certain success probability on an item. This is what are called the probability points of an item.

For instance, the p-60 point of an item is the ability level at which there is a probability of 0.60 of answering the item correctly. The p-points are easily determined. Consider the probability of correctly answering an item

$$p_i(\theta) = \frac{\exp(\alpha_i(\theta - \beta_i))}{1 + \exp(\alpha_i(\theta - \beta_i))}$$

For a given probability, the ability pertaining to that point is then determined from

$$\ln \frac{p_i(\theta)}{1 - p_i(\theta)} = \alpha_i(\theta - \beta_i),$$

from which it follows that

$$\theta = \beta_i + \frac{1}{\alpha_i} \ln \frac{p_i(\theta)}{1 - p_i(\theta)}.$$

Then the p-x point (with a probability of x) of an item is defined as

$$(p-x)_i = \beta_i + \frac{1}{\alpha_i} \ln \frac{x}{(1 - x)}.$$

It is easily seen that the p-50 point of an item equals the difficulty parameter  $\beta_i$ . If the item selection in a CAT takes places on the basis of success probability, this can be achieved as follows. Select the item for which the distance between the current ability estimate and the  $(p-x)_i$  point is minimal:

$$\min_i | \hat{\theta} - (p-x)_i |.$$

### 7.3.1 Performance of item selection based on nearest p-point

Simulation studies were conducted to evaluate the performance of the item selection methods. First, the results of a simulation study with an item bank calibrated with the 1pl will be given, followed by a study with an item bank calibrated with the 2pl.

The one-parameter model item bank

The 1pl item bank consists of 300 items with  $\beta \sim N(0,1)$ . The CAT algorithm used starts with an item of intermediate difficulty (one item randomly selected from 114 items with  $-0.5 < \beta_i < 0.5$ ) and has a fixed test length of 40 items. In the simulation, samples of 4000 abilities were drawn from the normal distribution:  $\theta \sim N(0,1)$ . Because of its profitable statistical properties, the weighted maximum likelihood estimator (Warm, 1989) was used for the estimation of the abilities. The selection methods at the different success probabilities were compared. As baselines in the comparison, the simulations were also conducted with random selection of all items and the optimal maximum information selection at the current ability estimate. The results of the simulations are given in Table 7.1.

Table 7.1: simulation 1pl CAT: selection nearest p-point.

Selection method	Mean error $1/n \sum_i (\hat{\theta}_i - \theta_i)$	mean se ( $\hat{\theta}_i$ ) (sd)	mean % correct (sd)
Max info	0.006	0.328 (0.015)	49.7 ( 8.6)
P_10	0.041	0.435 (0.009)	22.4 (14.5)
P_20	0.048	0.384 (0.045)	27.3 (12.4)
P_30	0.035	0.352 (0.024)	33.5 (10.3)
P_40	0.016	0.334 (0.015)	41.1 ( 8.8)
P_50	-0.013	0.328 (0.012)	50.0 ( 8.5)
P_60	-0.016	0.333 (0.017)	58.1 ( 9.2)
P_70	-0.029	0.351 (0.024)	65.4 (11.0)
P_80	-0.043	0.379 (0.044)	71.4 (13.5)
P_90	-0.034	0.424 (0.098)	75.2 (15.6)
Random	0.007	0.383 (0.078)	50.0 (19.9)

First it should be noted (second column of Table 7.1) that there is, on average, a small discrepancy between the known abilities and the estimated abilities, and



the effect seems to be systematic: when the items with a success probability lower than 0.50 are chosen, there is an overestimation of the mean ability; when selection takes places with higher success probabilities, the ability is generally slightly underestimated. This effect is in line with the known small bias of the ability estimator used (Warm, 1989) and is opposite to the bias in the maximum likelihood estimator of the ability as was reported in Pitkin & Vispoel (2001) when a test is not optimally assembled at an ability level. In section 7.4.2, we take a closer look to the remaining small bias in the ability estimates.

The selection methods show an effect in the desired direction in the results on the percentages correct (column 4 of Table 7.1). Selecting at a success probability higher or lower than 0.50 does not necessarily lead to the same percentage of correct answers of the simulated examinees. The more extreme the probability is, the larger the discrepancy between the selection percentage and the percentage correct. This can be explained by the fact that only a limited number of extremely difficult and extremely easy items are available in the item bank. (See also section 7.4.2.)

If we look in column 3 in Table 7.1 at the mean of the standard errors of the ability estimates with the selection methods, the expected effect can be seen. In the 1pl model, selection at maximum information is equivalent to selection of the item at the nearest p-50 point. Non-optimal selection, at other success probabilities, has an expected negative effect on measurement precision. The effect with the current item bank is symmetric around the p-50 point selection: selection at the nearest p-(50+x) point leads to about the same loss in precision as selection at the nearest p-(50-x) point.

The performance of selection methods can be compared more easily if the mean of the standard errors are considered as a function of the test length. The results for the selection methods with a success probability of higher than 50% are plotted in Figure 7.3

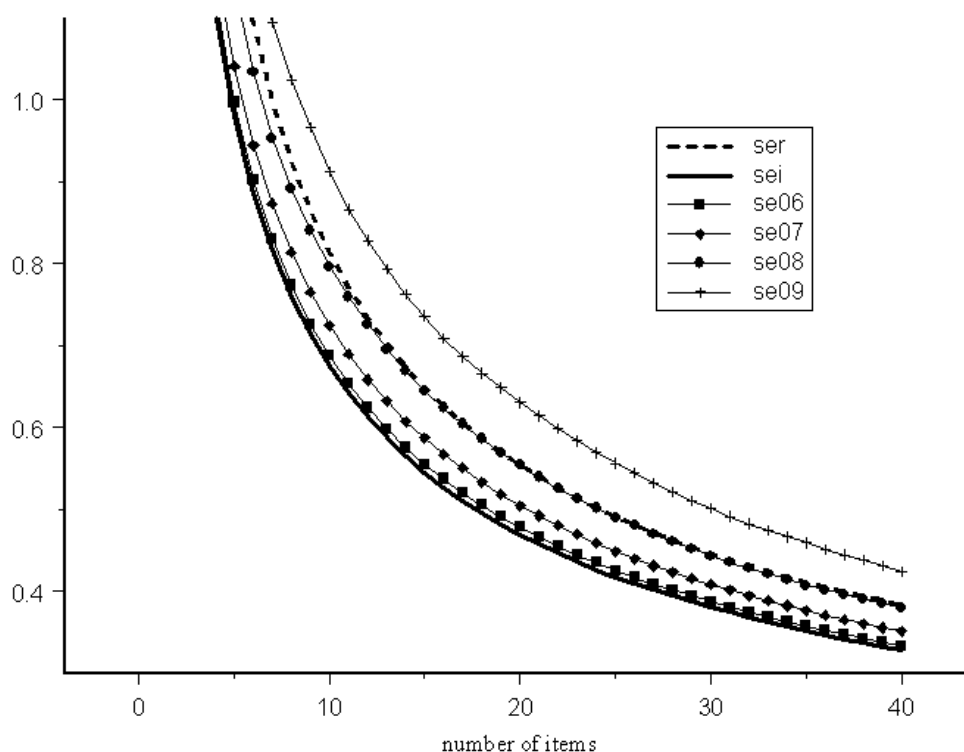


Figure 7.3. Mean se ability estimates and test length; nearest p-point selection; 1pl

It can be seen that the easier the selected items, the greater the loss in measurement precision is. The loss in precision when items are selected at the nearest p-60 and p-70 point is rather small. Selecting at p-80 is as bad as random selection, while selection at p-90 is even worse. Because in the 1pl model selecting at the nearest p-50 point is equivalent to maximum information selection, only maximum information is in Figure 7.3 (sei in the legend).

Table 7.2 gives the number of items needed on average with a selection method to achieve measurement precision which is equivalent with a test of 30 randomly drawn items from the bank.

Table 7.2. 1pl bank; selection on nearest p-points; equivalence with 30 random items

Selection method	Number of items
Max info	22
p-60	23
p-70	25
p-80	30
p-90	37

### The two-parameter model item bank

The 2pl item bank consists of 300 items with  $\beta \sim N(0,0.35)$  and  $\ln \alpha \sim N(0,0.35)$ . The CAT algorithm used starts with an item of intermediate difficulty (one item randomly selected from 113 items with  $-0.17 \leq \beta \leq 0.17$ ) and has a fixed test length of 40 items. In the simulation, samples of 4000 abilities were drawn from the normal distribution:  $\theta \sim N(0,0.35)$ . The selection methods at the different success probabilities are compared in Table 7.3.

Table 7.3: simulation 2pl CAT: selection nearest p-point.

Selection method	Mean error $1/n \sum_i (\hat{\theta}_i - \theta_i)$	mean se ( $\hat{\theta}$ ) (sd)	mean % correct (sd)
Max info	0.001	0.085 (0.013)	49.0 (11.9)
P_10	0.011	0.117 (0.033)	26.5 (18.5)
P_20	0.008	0.116 (0.020)	28.5 (14.1)
P_30	0.005	0.114 (0.013)	34.1 (10.6)
P_40	0.001	0.110 (0.010)	41.6 ( 8.9)
P_50	0.001	0.111 (0.008)	49.7 ( 8.5)
P_60	-0.009	0.110 (0.009)	58.0 ( 9.4)
P_70	-0.006	0.115 (0.015)	64.9 (11.8)
P_80	-0.009	0.114 (0.018)	70.8 (15.0)
P_90	-0.012	0.124 (0.033)	74.5 (17.6)
Random	0.001	0.132 (0.033)	49.7 (19.5)

The results for the mean percentages correct are about the same as in the case of the 1pl item bank (see Table 7.1). The same is true for the sign of the small bias in the ability estimates. The results on measurement precision show that selecting on higher or lower p-points has a very negative impact compared to maximum information selection. One sees that the more extreme the success probabilities are, the larger the loss in precision is, but in any case the loss is considerable, which is even clearer from Figure 7.4.

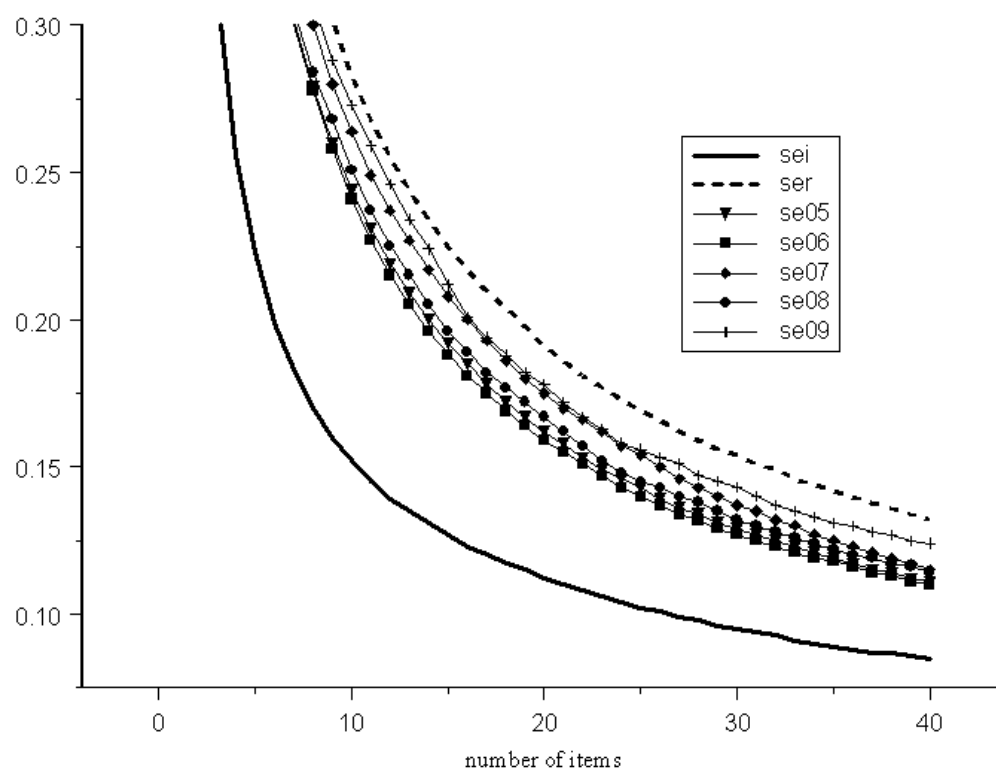


Figure 7.4. Mean se ability estimates and test length; nearest p-point selection; 2pl

Table 7.4 gives the number of items needed on average with a selection method to get measurement precision which is equivalent to a test of 30 randomly drawn items.

Table 7.4. 2pl bank; selection on p-points; equivalence with random test of 30 items

Selection method	Number of items
Max info	10
p-50	22
p-60	21
p-70	25
p-80	23
p-90	26

It is seen that selecting at the nearest p-60 point doubles the number of items needed compared to maximum information selection.

On the basis of the results presented in this section, it can be concluded that selecting at the nearest p-point of an item works quite well in an item bank based on the 1pl model, but for item banks calibrated with the 2pl model, the results are very poor. An explanation for this will be given in section 7.4 and an alternative selection method will be presented.

## 7.4 Alternative method for selecting with higher or lower success probabilities

The problem encountered with selection on success probability is due to the fact that, in selection, only the success probability of an item is considered, but not the values of the information function of the items. In the 1pl model, this has no consequences owing to the fact that, in that case, all information functions have the same shape; they only differ in the point where they reach their maximum ( $\theta = \beta_i$ ). This implies that the differences between, for instance, a p-50 point and a p-60 point of the item is constant for every item. In the 2pl model, however, the value of the information function plays an important role. Neglecting this and selecting only on the basis of success probability has negative consequences.

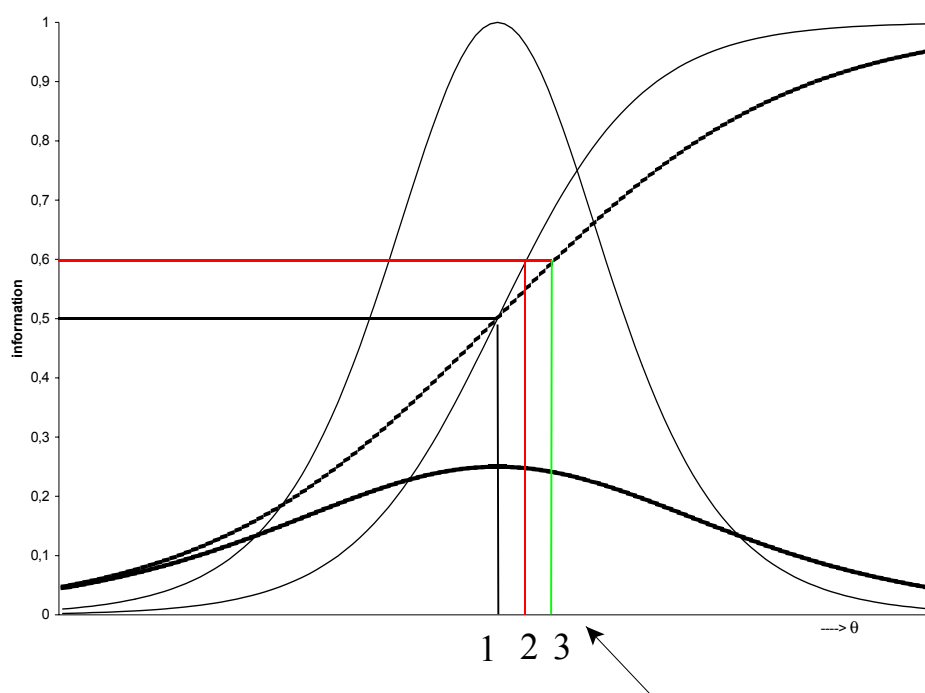


Figure 7.5. Item response functions and information functions of two items

What goes wrong is illustrated in Figure 7.5, which shows the item response curves and the information functions of two items with the same difficulty but with different discrimination parameters. For the first item, the discrimination parameter is  $\alpha_i = 1$  (dotted curves); for the second item,  $\alpha_i = 2$ . The coinciding p-50 point of both items is given on the ability axis at point 1; at point 2 and 3, the p-60 point of item 2 and item 1 respectively. If we select items at the nearest p-50 point, we can see that if the current ability estimate is at point 1, for this method, both items could be chosen, while at this point the information in item 2 is much higher. Another example: if we select at the nearest p-60 point and the current ability estimate is at the indicated arrow or higher, item 1 is preferred, while the value of the information function is much higher for item 2.

In order to overcome this problem, a new selection method was developed which takes account of the success probability and of the value of the information function. The idea is not selecting items with maximum information

at the current ability estimate, but selecting the item with maximum information at a lower or a higher level of ability than the current ability estimate. If easier items (with higher success probabilities) are wanted, one chooses items which are optimal (have maximum information) at an ability level which is below the current ability estimate. If more difficult items are desired, the items are selected at an ability point above the current estimate.

Suppose the current ability estimate is  $\hat{\theta}$ . Then easier or harder items are selected by searching at an ability level of  $y - \hat{\theta}$ , with  $y$  positive for easier items and negative for harder items. The value of the shift on the ability can be deduced from the desired success probability. In the 2pl model, it yields

$$p_i(\theta) = \frac{\exp(\alpha_i(\theta + y - \beta_i))}{1 + \exp(\alpha_i(\theta + y - \beta_i))}.$$

From which it follows that

$$\alpha_i(\theta + y - \beta_i) = \ln \frac{p_i(\theta)}{(1 - p_i(\theta))}.$$

In order to get a certain success probability, the shift on the scale is

$$y = \frac{1}{\alpha_i} \ln \frac{p_i(\theta)}{(1 - p_i(\theta))}.$$

So, e.g., for selecting items with a desired success probability of 60%, items are selected which have maximum information at

$$\theta = \hat{\theta} - \frac{1}{\alpha_i} \ln 1.5.$$

In the one-parameter model, the selection at the shifted ability level method is equivalent to selecting items at the p-points nearest to the current ability estimate. In the two-parameter model, however, the selection is quite different, which will become clear in the evaluation in the next section.

### 7.4.1 Performance of item selection based on selection at a shifted ability level

The selection at the shifted ability level was evaluated with the same simulation setup as in section 7.3.1. The 2pl item bank simulation results are given in Table 7.5.

Table 7.5: simulation 2pl CAT: selection at shifted ability level

Selection method	Mean error $1/n \sum_i (\hat{\theta}_i - \theta_i)$	mean se( $\hat{\theta}$ ) (sd)	mean % correct (sd)
Max info	0.001	0.085 (0.011)	49.0 (12.2)
P_10	0.010	0.100 (0.018)	28.6 (16.1)
P_20	0.006	0.091 (0.012)	33.5 (14.2)
P_30	0.004	0.088 (0.012)	38.8 (13.4)
P_40	0.001	0.086 (0.013)	43.8 (12.6)
P_50	-0.003	0.085 (0.013)	49.4 (12.3)
P_60	-0.004	0.085 (0.012)	55.1 (12.2)
P_70	-0.002	0.088 (0.012)	60.9 (12.6)
P_80	-0.008	0.092 (0.015)	65.4 (14.1)
P_90	-0.011	0.101 (0.015)	71.7 (15.6)
Random	0.003	0.133 (0.031)	50.5 (19.6)

The results for the mean % correct are about the same as with selecting on nearest distance to p-points. (Compare to Table 7.3). The same is true for the systematic bias in the ability estimates, although there is hardly any bias with the new selection method. (More details on the bias are given in section 7.4.1.) The results on measurement precision show that selecting easier or harder items is possible with the new selection method without a large loss in precision. This result is also seen from Figure 7.6.



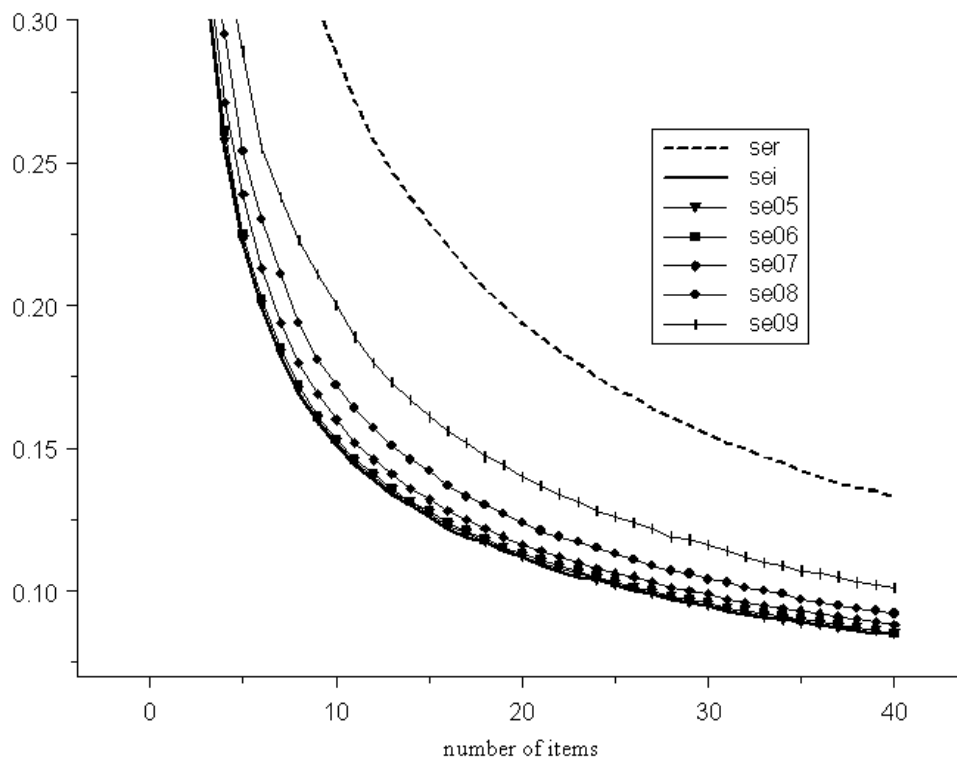


Figure 7.6. Mean se ability estimates and test length; selection at shifted ability; 2pl

It is clear that selecting easier or harder items with the new selection method does not cause much loss in measurement precision. If one aims at a success probability of 60%, there is hardly any loss: the more extreme the items are chosen, the larger the loss in efficiency. But at all success probabilities, the random selection is far outperformed in contrast to the results with the selection on the p-points. This result will become clearer in Table 7.6, which gives the average number of items needed with a selection method to get a measurement precision which is equivalent to a test of 30 randomly drawn items from the bank.

Table 7.6. 2pl bank; selection shifted ability level; equivalence with random 30 items

Selection method	Number of items
Max info	10
p-50	10
p-60	10
p-70	11
p-80	12
p-90	16

The new selection method seems to perform without any significant loss in measurement precision: with the current item bank and algorithm, it is possible to reach a percentage correct of 70% at the cost of, on average, 1 item compared to the optimal test.

#### ***7.4.2 Some properties of selection at the shifted ability level***

Three points were considered in more detail for the selection at the shifted ability. The bias of the ability estimate, the application of exposure control in the algorithm, and the effect of using a large item bank were explored.

##### The bias in the ability estimates.

In the evaluation of the selection at the shifted ability level, it was shown that the estimation error in the population,  $\theta \sim N(0, 0.35)$ , was almost zero. To explore the bias at distinct levels of the ability continuum, the simulations were also conducted at distinct values of  $\theta$ . For each selection method, 400 simulees were selected at 21 equidistant points between -1 and 1. The estimated abilities for the selection at the shifted ability aiming at 70% success probability are shown in Figure 7.7.

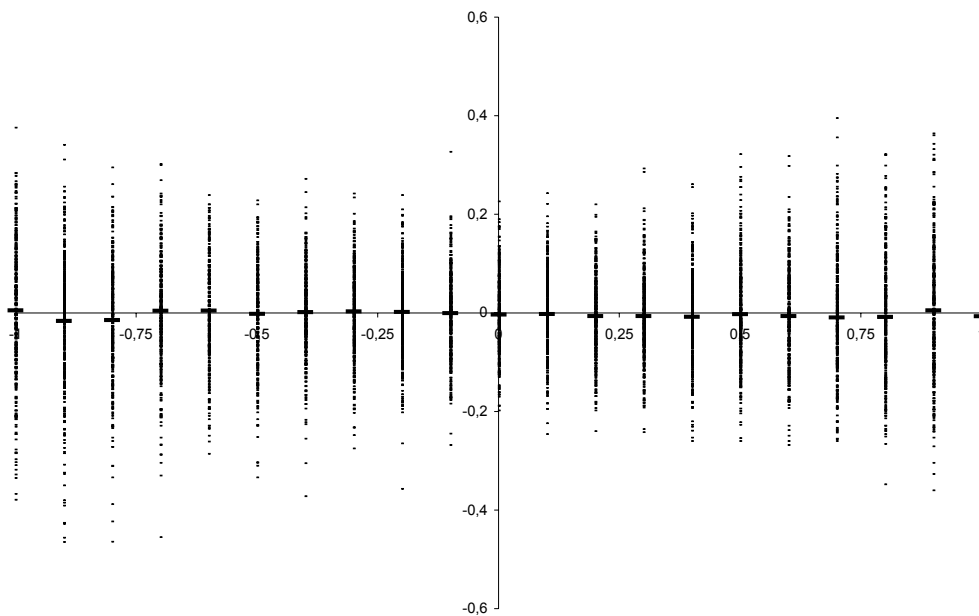


Figure 7.7. Ability estimates and ability; selection at shifted ability  $p=70$ ;  $2pl$

From this result it is clear that the variation in the ability estimates is about equal for all ability levels. This means that the small bias in the population reported in Table 7.5 is uniform for all ability levels. The simulations in which the selection took place at other  $p$ -levels yield the same results that were given for  $p=70$  in Figure 7.7.

#### Simulation with item selection applying exposure control

Because some form of exposure control is usually applied in the selection algorithm in modern CATs, it was investigated whether the new algorithm still works when exposure control is added to the CAT algorithm. The results of the same simulations, but combined with the application of the Simpson-Hetter exposure control with an maximum exposure of 0.3 (see Eggen, 2001), are given in Tables 7.7 and 7.8.

Table 7.7: simulation 2pl CAT: selection at shifted ability level and exposure control

Selection with SH 0.3	Mean error $1 / n \sum_i (\hat{\theta}_i - \theta_i)$	Mean se ( $\hat{\theta}$ ) (sd)	mean % correct (sd)
Max info	0.002	0.098 (0.010)	49.0 (10.0)
P_50	0.001	0.098 (0.008)	49.5 (10.0)
P_60	0.002	0.100 (0.011)	54.5 (11.4)
P_70	-0.017	0.104 (0.013)	55.5 (12.8)
P_80	-0.008	0.106 (0.011)	59.3 (14.4)
P_90	-0.005	0.111 (0.016)	60.0 (15.6)
Random	0.003	0.133 (0.031)	50.5 (19.6)

Again, there is hardly any bias and the differences in the percentages correct seem to be less than in selecting without exposure control. The discrepancy between the desired and the achieved percentages correct is larger when exposure control is applied. (Compare column 4 of Table 7.7 with the same column in Table 7.5). With respect to measurement precision, the results are similar to selecting without exposure control. The number of items needed to get an equivalent to a test with 30 randomly selected items is given in Table 7.8. It is clear that applying exposure control on average costs 2 or 3 items .

Table 7.8. 2pl bank; selection shifted ability level and exposure control; equivalence with 30 random items

Selection method with SH =0.3	Number of items
Max info	12
p-50	12
p-60	13
p-70	14
p-80	15
p-90	18

Simulations with a large item bank.

A possible explanation for this discrepancy between the desired and the achieved percentages correct is that there is a mismatch between the items available in the item bank and the desired percentages in the population. One possible solution for this could be enlarging the size of the item bank. In order to check this, simulations were conducted with a very large item bank.

The 2pl item bank consists of 3000 items, 1000 with  $\alpha=2$ ,  $\alpha=3$  and  $\alpha=4$  and the difficulty parameter from a uniform distribution  $\beta \sim U(-1.1, 1.1)$ . The CAT algorithm used starts with an item of intermediate difficulty and has a fixed test length of 40 items. In the simulation, samples of 4000 abilities were drawn from the normal distribution:  $\theta \sim N(0, 0.35)$ . The selection methods for different success probabilities are compared in Tables 7.9 and 7.10.

Table 7.9. Simulation 2pl CAT large item bank; selection at shifted ability level.

Selection method	Mean error $1 / n \sum_i (\hat{\theta}_i - \theta_i)$	mean se ( $\hat{\theta}$ ) (sd)	mean % correct (sd)
Max info	-0.001	0.083 (0.002)	50.1 (7.6)
P_10	0.053	0.144 (0.031)	10.9 (5.7)
P_20	0.026	0.106 (0.014)	20.1 (6.5)
P_30	0.015	0.091 (0.007)	30.2 (7.0)
P_40	0.005	0.085 (0.004)	39.6 (7.5)
P_50	0.000	0.083 (0.002)	50.0 (7.7)
P_60	-0.004	0.085 (0.004)	60.1 (7.7)
P_70	-0.015	0.091 (0.007)	70.2 (6.9)
P_80	-0.027	0.106 (0.013)	79.7 (6.4)
P_90	-0.056	0.143 (0.031)	88.9 (6.0)
Random	-0.003	0.145 (0.015)	50.4 (16.0)

We see here the same results as reported before, except that the desired percentages correct are now in line with the percentages that are achieved. Finally, the number of items needed for the large item bank to get a precision

equivalent to a test with 30 randomly selected items is given in Table 7.10. The results again show that selecting at a shifted ability level, up to the p-70 level, has only a limited loss in precision as a result.

*Table 7.10 Large item bank (2pl ); selection shifted ability level; equivalence with random test of 30 items*

Selection method	Number of items
Max info	11
p-50	11
p-60	12
p-70	13
p-80	17
p-90	29

## 7.5. Discussion

In this study, it was shown that, in CATs, it is possible to select items with a higher or lower success probability. The selection methods based on the minimal distance between the current ability estimate and the p-points of the items works only satisfactorily if the item bank is calibrated with the 1pl model. This selection method yields unsatisfactory results when it is applied to an item bank which is calibrated with the 2pl model.

The method introduced, in which items are chosen that have maximum information at an ability level lower or higher than the current ability estimate, also performs well in item banks calibrated with the 2pl model. With item banks of a practical size (300), a little loss in measurement precision is the price of a (somewhat) easier or more difficult test. The method is also effective if the selection is combined with the application of exposure control. Getting very high or very low percentages correct was seen to be possible with a larger item bank. In that case, in principle, any desired percentage correct could be reached, but extreme values of the success probabilities are combined with a considerable loss

in precision. For practical purposes, item selection, aiming at percentages correct of 60 or 70 (or 40 or 30), seems to be possible without a large loss in precision.

It can be mentioned that all the selection methods and the results are symmetric around the p-50 points. For the selection methods, this is only true for the 1pl and 2pl model. Knowing that the symmetry disappears, it is worthwhile investigating the application of the selection method if the 3pl model, including a guessing parameter, is used.

Finally, it should be noted that knowing the effect of selecting with other success probabilities in mind, one could, for CAT applications, build item banks which are more suitable for that purpose. The item banks studied here are in a sense optimal for a CAT with maximum information item selection: the mean difficulty of the items is equal to the mean of the population. If one knows, for instance, that one wants an easy CAT, one could try to construct a bank which is, on average, easier for the population.

## 7.6 References

- Bergstrom, B.A., Lunz, M.E., & Gershon, R.C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education*, 5, 137-149.
- Eggen, T.J.H.M. (2001). *Overexposure and underexposure of items in computerized adaptive testing*. Measurement and Research Department Reports, 2001-1. Arnhem: Cito.
- Eggen, T.J.H.M., & Straetmans, G.J.J.M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological measurement*, 60 , 713-734
- Hambleton, R.K., & Swaminathan, H. (1985) *Item response theory: Principles and applications*. Boston: Kluwer Academic Publishers
- Kingsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4, 241-261.
- Pitkin, A. K., & Vispoel, W.P.(2001) Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement*, 38, 235-247.
- Rocklin, T.R., & O' Donnell, A.M., (1987). Self-adapted testing: A performance improving variant of computerized adaptive testing. *Journal of Educational Psychology*, 79, 315-319.
- Van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. (pp. 1-25) In: W.J. vand Linden, & C.A.W. Glas. (Eds.). *Computerized adaptive testing. Theory and practice*. Dordrecht: Kluwer Academic Publishers
- Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A Primer. Second Edition*. Hillsdale (NJ): Lawrence Erlbaum.
- Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-450.





corresponding to the midpoint of the critical inequality interval which is closest to the current examinee's score. If the 2-PL is the IRT model, the midpoints follow from (19) and (20); after determining the minimum of  $|\sum_i a_i x_i - (U_1 + L_1)/2|$  and  $|\sum_i a_i x_i - (U_2 + L_2)/2|$ , the item with maximum information at the corresponding cutting point will be selected.

### 6.3.2 Kullback-Leibler information

The item selection methods described in the preceding section all use a criterion related to Fisher information that has a strong relation to optimizing estimates. Although these selection methods can also be used in adaptive testing with the SPRT, one could wonder whether making use of Fisher information is optimal in this case. An alternative could be to base the item selection process on the relative entropy or Kullback-Leibler information (Cover & Thomas, 1991), which is an information concept as strongly related to statistical testing as Fisher information is to statistical estimation. The relative entropy is a measure of the discrepancy between two distributions:

$$K(f_1 || f_0) = \mathcal{E}_{f_1} \log \left( \frac{f_1(x)}{f_0(x)} \right), \quad (21)$$

which is the expected information in  $X$  for discrimination between the two hypotheses  $H_0: f(x) = f_0(x)$  and  $H_1: f(x) = f_1(x)$ . The larger this information, the more efficient the statistical test will be. (The notation with the double vertical bars is standard for the K-L distance between  $f_1$  and  $f_0$ ).

The definition in (21) can be directly applied to the SPRT application in adaptive testing:  $H_0$  is the hypothesis that we have a distribution (likelihood) with parameter value  $\theta = \theta_1$  and under  $H_1$  the distribution has parameter  $\theta = \theta_2$ . And

$$K(\theta_2 || \theta_1) = \mathcal{E}_{\theta_2} \log \left( \frac{L_k(\theta_2; \mathbf{x})}{L_k(\theta_1; \mathbf{x})} \right) \quad (22)$$

is the Kullback-Leibler test information ( $k$  items), which can be written as the

sum of the Kullback-Leibler information of the items:

$$K(\theta_2 || \theta_1) = \sum_{i=1}^k K_i(\theta_2 || \theta_1) = \sum_{i=1}^k \mathcal{E}_{\theta_2} \log \left( \frac{L(\theta_2; x_i)}{L(\theta_1; x_i)} \right) \quad (23)$$

The K-L item information  $K_i(\theta_2 || \theta_1)$  is defined for any pair  $\theta_2$  and  $\theta_1$  and is a positive real number and, consequently, an eligible item information index. The usefulness of applying an item selection procedure based on maximum K-L information can be understood, since this procedure will maximize the contribution to the K-L test information. When the K-L test information, (23), is maximized, the expected difference between the log likelihoods under both hypotheses is maximized. Which is the same as making the likelihood ratio more extreme, which is, in turn, expected to minimize the number of items needed to take a decision because the test statistic is the likelihood ratio (see (5)).

K-L information is also the basis for an index proposed by Chang and Ying (1996) for estimation problems as a more global information index in contrast to the local Fisher information. They consider, for any  $\theta$ , the K-L item information to the true ability  $\theta_0$ :  $K_i(\theta_0 || \theta)$ , which is then, of course, a function of  $\theta$ . They define their information index which is used in item selection as an integral of this function over an interval depending on the current MLE,  $\hat{\theta}_k$ , and an expression,  $\delta_k$ , which is decreasing in the number of items ( $k$ ):

$$K_i(\hat{\theta}_k) = \int_{\hat{\theta}_k - \delta_k}^{\hat{\theta}_k + \delta_k} K_i(\hat{\theta}_k || \theta) d\theta. \quad (24)$$

Chang and Ying's (1996) claim is that their information measure is a good alternative, especially in the beginning of the test, when the ability of an examinee is poorly estimated. It should be noted that information indices like the one given in (24), but then based on Fisher information, were also proposed by Veerkamp and Berger (1997). But, because these indices are not expected to be useful alternatives for item selection in the case of the SPRT, they will not be discussed further here.

If an IRT model for dichotomously scored items is used, the proposed K-L item

information index, (23), can be written as:

$$K_i(\theta_2 || \theta_1) = p_i(\theta_2) \log \frac{p_i(\theta_2)}{p_i(\theta_1)} + q_i(\theta_2) \log \frac{q_i(\theta_2)}{q_i(\theta_1)}, \quad (25)$$

which with the 2-PL model specializes to:

$$K_i(\theta_2 || \theta_1) = a_i(\theta_2 - \theta_1) p_i(\theta_2) + \ln \frac{q_i(\theta_2)}{q_i(\theta_1)}. \quad (26)$$

Note that the index is linear in the discrimination parameter, whereas the Fisher item information is quadratic in  $a_i$ , which means that the weight of the discrimination parameter in the selection is less, which can be favorable in the beginning of the test.

If the K-L information is computed in  $\theta_2 = \theta_0 + \delta$  and  $\theta_1 = \theta_0 - \delta$ , equation (26) becomes

$$K_i(\theta_0 + \delta || \theta_0 - \delta) = \frac{2a_i\delta \exp a_i(\theta_0 + \delta - b_i)}{1 + \exp a_i(\theta_0 + \delta - b_i)} + \ln \left[ \frac{1 + \exp a_i(\theta_0 - \delta - b_i)}{1 + \exp a_i(\theta_0 + \delta - b_i)} \right], \quad (27)$$

which is a monotone increasing function of  $\delta$ . This means that for any fixed item  $i$ , the K-L item information increases if the width of the indifference zones increases. This property illustrates that the K-L item information expresses the contribution of an item to the capability to distinguish between two hypotheses, which is larger when  $\delta$  is larger. However, this property does not imply that the order of the item K-L information over items is the same for each  $\delta$ .

#### Some selection procedures based on K-L information.

K1 In the case of a classification problem in two categories, the K-L item information can be used directly in a straightforward way in item selection. The K-L item information will be computed in two points symmetric around the cutting point:

Select the item  $i$  for which:  $\max_i K_i(\theta_0 + \delta || \theta_0 - \delta)$ .

K2 In the three-way classification<sup>i</sup> for K-L item selection, there are more

possibilities. One possibility is to select the item which maximizes the K-L information at two fixed points. Possible choices are (see Figure 6.1): a.  $K_i(\theta_{21}||\theta_{12})$ , b.  $K_i(\theta_2||\theta_1)$  and c.  $K_i(\theta_{22}||\theta_{11})$ , which have in common that the items will be selected with maximum information to distinguish between two hypotheses. This may cause a problem, because a decision in one of three categories is needed.

- K3 One way to deal with this problem is, as with Fisher information (see F3 before), is to look for the nearest cutting point and to select the items with maximum K-L information around this cutting point. The nearest cutting point is, as in F4, determined without estimation by comparison of the score with the midpoints of the critical intervals of the tests.
- K4 An alternative approach is to look more precisely at the progress of hypothesis testing: as long as none of the pairs of hypotheses have led to a decision, items are chosen with maximum K-L information between the two cutting points  $\theta_1$  and  $\theta_2$ ; if one of the pairs of hypotheses has led to a decision while the other has not, items will be chosen which have maximum K-L information around the cutting point corresponding to the test which has not yet led to a decision. If the 2-PL model is used, this selection procedure can be described as follows. Select the item  $i$  for which:

$$\text{if: } \sum_{i=1}^k a_i x_i \geq U_1: \quad \max_i K_i(\theta_{22}||\theta_{21}) \quad (28)$$

$$\text{if: } \sum_{i=1}^k a_i x_i \leq L_2: \quad \max_i K_i(\theta_{12}||\theta_{11}) \quad (29)$$

$$\text{else: } \max_i K_i(\theta_2||\theta_1). \quad (30)$$

A variation on this procedure could be made in case no decision has been taken yet: instead of the expression in (30), a narrower interval is chosen:

$$\max_i K_i(\theta_{21}||\theta_{12}). \quad (31)$$

## 6.4 Comparison of item selection procedures

For both the two-category and the three-category classification problem the performance of the item selection procedures described in the preceding section were investigated.

### 6.4.1 Method

The performance of the item selection procedures were evaluated and compared to each other by means of simulation studies. For the simulation studies an operational item bank was used. This item bank contains 250 mathematics items which are used in adult education to place students in one of three course levels and to measure the progress at these levels. Most of the items have an open-ended short answer format, but all the items are scored dichotomously. The items in the bank belong to one of three content subdomains: mental arithmetic/estimating, measuring/geometry and the other elements of the curriculum. Despite these subdomains, the items were shown to fit the one-dimensional 2-PL model. The scale was fixed by restrictions on the item parameters. The mean item difficulty is 0, and the mean discrimination is 3.09. On this scale, the distribution of the ability  $\theta$  in the population was estimated to be normal with a mean of .294 and a standard deviation of .522. More details on the scaling can be found in Eggen and Straetmans (2000).

The simulations were conducted as follows. An ability of a simulee  $\theta_v$  was randomly drawn from  $N(.294, .522)$ . Three relatively easy starting items were selected; subsequent items were selected using one of the investigated item selection methods. The simulee's response to an item was generated according to the IRT model and this procedure was repeated for  $N=5000$  simulees. For varying decision error rates, the item selection procedures were compared on the mean number of items required to make a decision and the classification accuracy, the percentages of correct decisions.

For the classification in two categories, the cutting point on the ability scale in the simulations was  $\theta_0=.1$ , and the maximum test length was:  $k_{\max}=25$ . The adaptive testing procedures were conducted for three different error rates:  $\alpha = \beta$

were .05, .075 and .1 and varying indifference zone:  $.1 \leq \delta \leq .23$ , in steps of .01.

The following item selection procedures were compared:

- F1. Maximum Fisher information at the current estimate.
- F2. Maximum Fisher information at the cutting point.
- K1a. Maximum K-L information at  $\theta_1 = .05$  and  $\theta_2 = .15$ .
- K1b. Maximum K-L information at  $\theta_1 = .00$  and  $\theta_2 = .20$ .
- K1c. Maximum K-L information at  $\theta_1 = -.05$  and  $\theta_2 = .25$ .

For the classification problem in three categories, the cutting points in the simulation were  $\theta_0 = -.13$  and  $\theta_0 = .33$ , and the maximum test length was:  $k_{\max} = 25$ . The adaptive testing procedures were conducted for three different sets of error rates:  $\alpha_1 = \beta_2 = 2\beta_1 = 2\alpha_2$  were .05, .075, and .1. Halving  $\beta_1$  and  $\alpha_2$  compared to  $\beta_2$  and  $\alpha_1$  has the effect that it is expected that all three decisions will have the same error rate. The width of the indifference zones was also varied:  $.10 \leq \delta \leq .20$ , in steps of .01. No  $\delta$  larger than .2 was considered, as the zones of both hypotheses would then overlap.

The following item selection procedures were compared:

- F1. Maximum Fisher information at the current estimate.
- F3. Maximum Fisher information at the cutting point nearest to the current estimate.
- F4. Maximum Fisher information at the nearest cutting point.
- K2a. Maximum K-L information at  $\theta_1 = -0.03$  and  $\theta_2 = 0.23$ .
- K2b. Maximum K-L information at  $\theta_1 = -0.13$  and  $\theta_2 = 0.33$ .
- K2c. Maximum K-L information at  $\theta_1 = -0.23$  and  $\theta_2 = 0.43$ .
- K3. Maximum K-L information at the nearest cutting point.
- K4a. Maximum K-L information at varying points: see (28), (29), and (30).
- K4b. Maximum K-L information at varying points: see (28), (29), and (31).

In the use of the mathematics item bank for the classification problem, there is a strong practical wish that every adaptive test is in agreement with certain content specifications. The specifications are that a test would preferably consist of 16% items from the first subdomain (arithmetic/ estimating), 20% items from the second domain (measuring/ geometry) and 64% items from the third

subdomain. In testing algorithms this constraint on an item selection procedures can be achieved by applying the Kingsbury and Zara (1991) approach: after each administered item the difference between the desired and the achieved percentages of items selected from each subdomain is calculated. And then from the domain for which this difference is largest, the next item is chosen according one of the maximum information selection procedures. In order to explore the effect of imposing the content constraint to the item selection on the possible differences between the item selection procedures, this was investigated for three of the above mentioned selection procedures.

### 6.4.2 Results

For the classification problem in two categories, the results for the indifference zone with  $\delta=.15$  are given in Table 6.1.

*Table 6.1. Mean number of required items and percentage of correct decisions in a decision problem with one cutting point  $\theta_0=.1$  and with  $\delta=.15$ .*

	selection method									
	F1		F2		K1a		K1b		K1c	
	k	%	k	%	k	%	k	%	k	%
error rate										
$\alpha = \beta = .05$	16.0	95.6	16.3	94.7	16.1	95.4	16.3	94.6	15.9	95.5
$\alpha = \beta = .075$	14.9	95.0	14.0	95.2	13.9	94.8	13.9	95.2	13.9	95.6
$\alpha = \beta = .10$	13.2	94.9	12.7	94.8	13.2	94.8	12.7	95.3	12.9	94.8

Most notable is that there are almost no differences in Table 6.1. For the three error rates and all five item selection methods, the percentage of correct decisions are about 95%. A consistent difference between these error rates over selection methods can be seen in the mean number of required items. For instance, with an error rate of .05 about 16 items are needed, and if the rate is .1 on average 3 items less are required. In general, it is seen that the lower the rates, the more items are needed. There are hardly any differences between the three variants of K-L information selection. There seems to be a slight tendency, at least when the error rates are .075 and .10, for the selection of items with maximum Fisher information at the cutting point (F2) to be better than the selection of items with



maximum Fisher information at the current estimate (F1). This is in line with the findings of Spray and Reckase (1996). Furthermore, K-L information selection (K1) seems to be as good as selecting with maximum Fisher information at the cutting point (F2).

These results are also found if the indifference zone  $\delta$  is varied. The results for  $\alpha = \beta = .1$  for the three selection procedures F1, F2 and K1 are given in Figures 6.3 and 6.4.

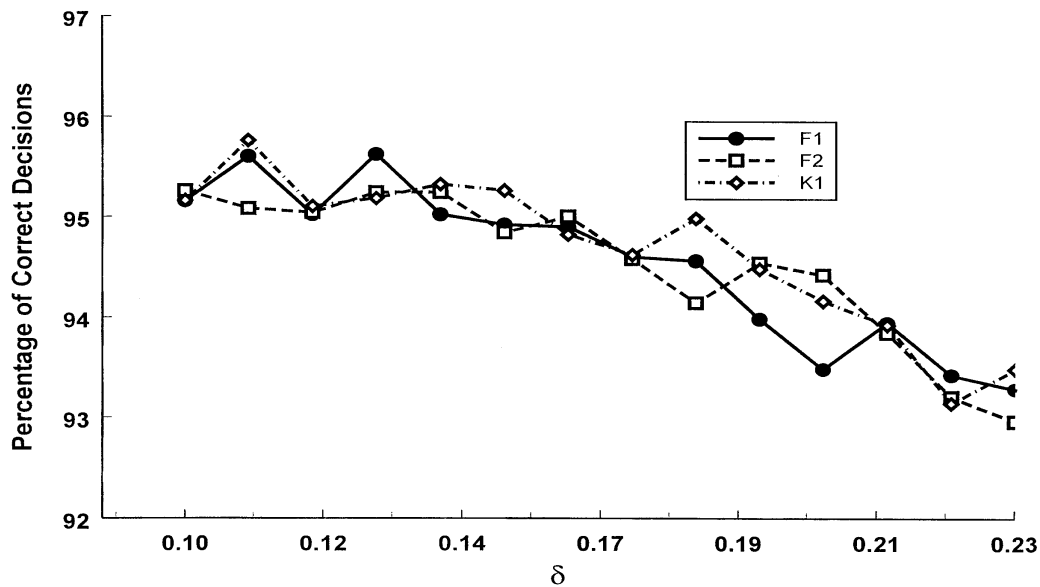


Figure 6.3. Percentage correct decisions for three item selection procedures in the two-category problem as a function of  $\delta$ .

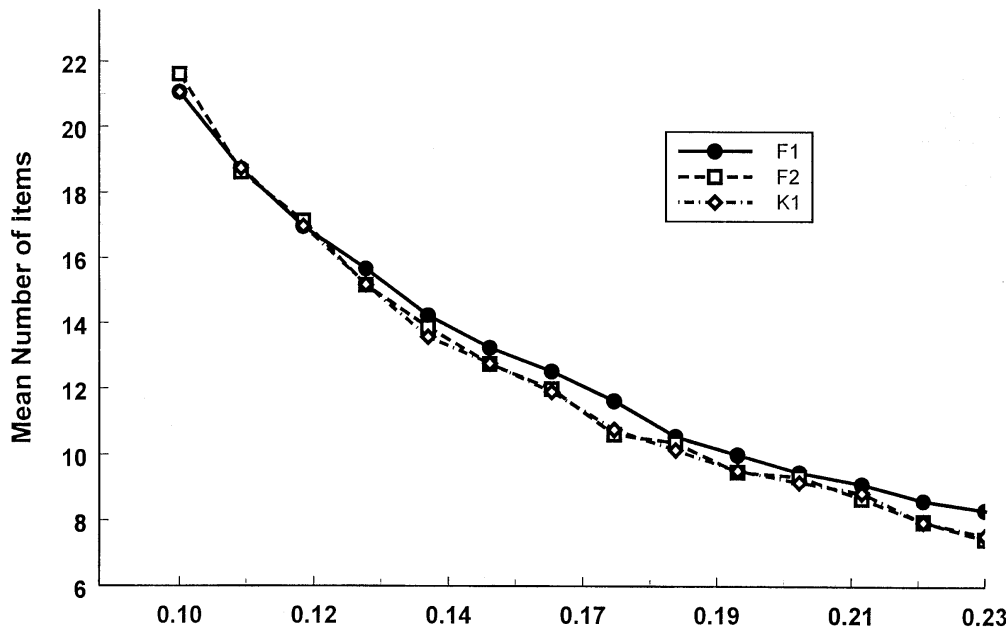


Figure 6.4. Mean number of required items for three item selection procedures in the two-category problem as a function of  $\delta$ .

In Figure 6.3, a slight decrease in the percentages of correct decisions with increasing  $\delta$  is seen. This can be understood because when the indifference interval is larger, for more ability levels the acceptable decision error rate is not guaranteed which will have a negative effect on the percentage of correct decisions. The effect is seen for all three item selection methods, which compared to each other show hardly any differences.

Figure 6.4 shows that, as expected, the mean number of items required decreases as the indifference zone increases for all three selection procedures. Furthermore, it can be seen that the K-L information item selection and maximum Fisher information selection at the cutting point (F2) is as good as and, from  $\delta > .12$ , a bit better than selection with Fisher information at the current estimate (F1).

For the classification problem in three categories the results for  $\delta = .13$  are given in Table 6.2. It is seen that for every selection method, there is an expected decrease in the mean number of required items if the acceptable error rates are increased. The differences hardly vary if the selection methods are compared. Doubling the acceptable error rate gives a decrease of on average about 2.5 items.

Increasing the error rates has little effect on the percentages of correct decisions.

Table 6.2. Mean number of required items and percentage of correct decisions in a decision problem with two cutting points  $\theta_1 = -.13$ ,  $\theta_2 = .33$  and with  $\delta = .13$ .

selection	error rates					
	.05		.075		.1	
	k	%	k	%	k	%
F1	16.7	89.9	15.6	89.2	14.6	89.1
F3	21.8	87.0	20.5	87.7	19.4	87.4
F4	16.8	89.6	15.6	90.0	14.3	88.5
K2a	18.7	89.5	17.4	89.5	16.3	88.1
K2b	18.4	88.4	17.0	88.0	16.3	88.6
K2c	18.7	87.9	17.1	88.2	16.4	88.6
K3	16.8	90.1	15.3	89.2	14.2	89.4
K4a	17.0	89.2	15.6	89.2	14.4	89.4
K4b	17.0	89.2	15.5	89.7	14.2	89.1

A comparison of the selection methods shows that the differences between them are consistent over the different error rates. Next, it is noted that with the K-L selection methods K2 and K4, varying the exact pair of points between which the K-L information is computed has no impact on the performance of the adaptive test. The same was already seen in the two-way classification problem. Obviously, varying  $\delta$  in the computation of the K-L item information has not a large impact on the ordering of the items on this information. In the three way classification problem, the worst performing selection method is clearly the one in which items are selected with maximum Fisher information at the cutting point nearest to the current estimate (F3). It needs more items and has a lower percentage of correct decisions. This finding, confirming those of Eggen and Straetmans (2000), may be explained by the fact that the current estimate of the ability, especially in the beginning of the test, is so inexact that it is sometimes nearer to the wrong cutting point than the cutting nearest to the true value of the ability. It is also clear that in decision problems with three categories, item selection which maximizes the K-L information at only two fixed points (K2) is

as could be expected worse than other methods. There seem to be four (F1, F4, K3, K4) selection methods in the three-category problem that perform almost equally well.

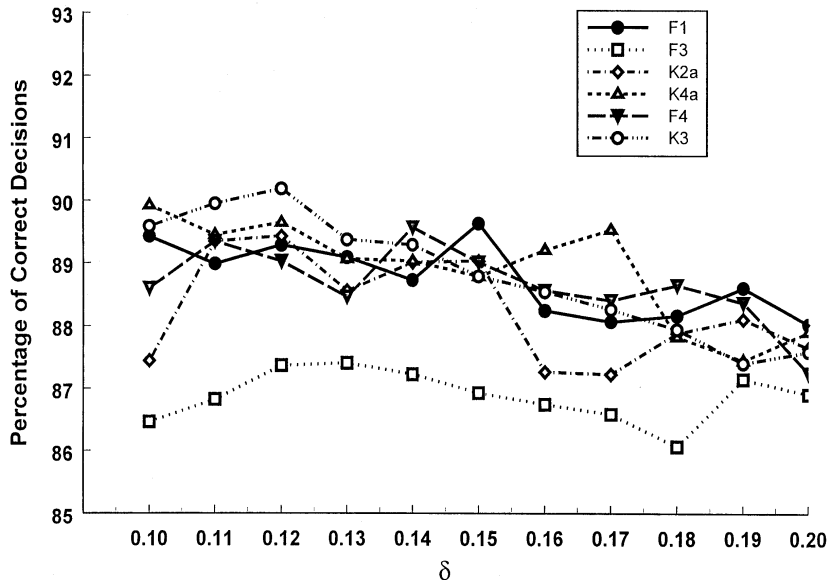


Figure 6.5. Percentage correct decisions for six item selection procedures in the three-category problem as a function of  $\delta$ .

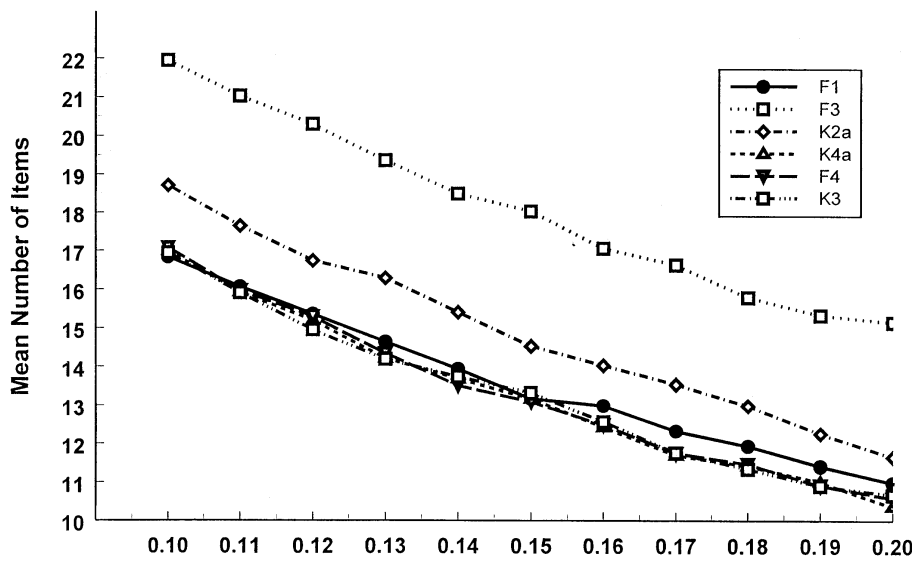


Figure 6.6. Mean number of required items for six item selection procedures in the three-category problem as a function of  $\delta$

In Figures 6.5 and 6.6, the simulation results are shown as a function of the indifference zone. The results mentioned before hold independent of the width of the indifference zone. In Figure 6.5, there are systematically lower percentages of correct decisions if the selection is based on maximizing the Fisher information at the cutting point nearest to the current estimate (F3). All the other selection methods perform about equal on the percentages of correct decisions. In Figure 6.6, the expected decrease in the mean number of required items with increasing  $\delta$  is again seen for all selection methods. On this aspect, the F3 method and the K2 methods (choosing items with maximum K-L information at two fixed points), clearly performed worse than the other four methods. Of these four methods, selecting items which maximum Fisher information at the current estimate (F1), with some indifference zones  $\delta$ , needs, on average, slightly more items than the K4, F3, and K3 methods. For instance for  $\delta=.16$ , F1 needs on average 12.98 items, and the other three respectively 12.42, 12.28 and 12.56 items. This could be a reason to prefer one of these three methods of item selection to selection on basis of Fisher information at the current estimate (F1). Of these three methods, F4 and K3 have in common the way the nearest cutting point is sought: the current weighted score (in the 2-PL) is compared with the midpoints of the critical interval of the two tests. Unfortunately, this is an ad hoc criterion which is not based on a clear concept and a generalization to the case the three parameter logistic model (3PL) including a guessing parameter is used as the IRT model, is not straightforward. Nevertheless, the performance of these selection methods in the 2PL case is as good as the conceptually better grounded and in general preferable K4 selection method.

The effect of content balancing on the performance of the item selection methods for three selection methods, which show fairly clear differences in Figure 6.6, is given in Table 6.3.

Table 6.3. Mean number of required items and percentage of correct decisions in a decision problem with two cutting points.  $\alpha_1 = \beta_2 = 2\beta_1 = 2\alpha_2 = .1$  and  $\delta = .13$ .

	selection method					
	F3		K2		K4	
	k	%	k	%	k	%
Content control						
no	19.4	87.4	16.3	86.1	14.2	89.1
yes	18.9	88.3	16.4	88.3	14.7	88.9

It is clear that with the mathematics item bank constraining the item selection methods with a content specification does not impair the quality of the procedures seriously. For all three studied item selection the differences are small. However, it is interesting to see that imposing the content control has a positive effect on the performance of the worst selection technique (F3), and a little negative effect on the methods based on the K-L information. This finding supports the expectation that differences between item selection methods will become smaller when more constraints are imposed to the selection methods.

## 6.5 Discussion

The results of the present study indicate that when the sequential probability ratio test is applied in adaptive testing, item selection methods can be defined which are based on an information concept which has a natural relation with hypothesis testing. These item selection methods are based on Kullback-Leibler information or relative entropy, which expresses the power of an item to discriminate between two hypotheses. For decision problems in two and three categories, item selection methods based on K-L information were given as an alternative for item selection methods which are based on the 'estimation-related' Fisher information.

The comparison of the performance of the item selection methods in the decision problem with two categories, showed there was no difference between maximizing K-L information around the cutting point and maximizing Fisher information at the cutting point, but both are slightly better than maximizing Fisher information at the current estimate of an examinee. In the decision problem with three categories, one of the best performing item selection methods was the

selection method which maximizes the K-L information between two varying hypotheses. The hypotheses to be considered depend on the progress of the testing thus far: if testing one of the two pairs of hypotheses has led to a decision, items are chosen with maximum information at the other pair of hypotheses; if none has reached a decision, the information is maximized between the two pairs of hypotheses.

The results reported in the paper are all based on the application of the 2PL model as the item response model. The generalization to the 3PL model as well as the SPRT testing procedure (see (9), (13) and (14)), as of the definition of the K-L item information is straightforward. The expressions look more complicated, but there are no principal differences. For instance, the K-L item information in  $\theta_2 = \theta_0 + \delta$  and  $\theta_1 = \theta_0 - \delta$ , the 3PL model analogue of equation 26, becomes

$$K_i(\theta_0 + \delta || \theta_0 - \delta) = \frac{c_i + \exp a_i(\theta_0 + \delta - b_i)}{1 + \exp a_i(\theta_0 + \delta - b_i)} \cdot \ln \frac{c_i + \exp a_i(\theta_0 + \delta - b_i)}{c_i + \exp a_i(\theta_0 - \delta - b_i)} + \ln \frac{1 + \exp a_i(\theta_0 - \delta - b_i)}{1 + \exp a_i(\theta_0 + \delta - b_i)}, \quad (32)$$

which is, as (26), monotone increasing in  $\delta$ , and monotone decreasing as a function of  $c_i$ , the guessing parameter. For applying the SPRT in the two-category decision problem, Spray and Reckase (1994) using the 3PL model report that item selection on the basis maximum Fisher information at the cutting point (F2) performs much better than selecting items with maximum information at the current estimate (F1). Whether the same loss of efficiency appears in the three-category problem and selecting with maximum Fisher information at the current estimate is not known. However there are no reasons to expect that the Kullback-Leibler based item selection method will have these problems in the 3PL case. On the contrary, in as well the two-category problem as the three-category the K-L item information selection resembles Fisher information selection at the cutting point more the Fisher information selection at the current estimate.

In SPRT adaptive testing item selection based on the conceptually strongly related K-L information is generally preferred to Fisher information-based

methods. In both the two- and three-category decision problem, the item selection based on K-L information never performed worse and sometimes better than Fisher information-based selection in the simulation study. Moreover, in some of the Fisher information-based item selection methods, an estimate of the current ability is needed. This is never the case in K-L information item selection which is computationally much easier.



## 6.6 References

- Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society, B*, 12, 137-144.
- Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, 20, 213-229.
- Cover, T.M. & Thomas, J.A. (1991). *Elements of information theory*. New York: Wiley.
- Eggen, T.J.H.M & Straetmans, G.J.J.M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, 66, 713-734.
- Kingsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4, 241-261.
- Lewis, C., & Sheenan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, 14, 376-386.
- Reckase, M.D. (1983). A procedure for decision making using tailored testing. In: D.J. Weiss (Ed.), *New horizons in testing* (pp. 237-255). New York: Academic Press.
- Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics*, 20, 502-522.
- Spray, J.A. (1993). *Multiple-category classification using a sequential probability ratio test*. (Research report 93-7). Iowa City: American College Testing.
- Spray, J.A., & Reckase, M.D. (1994). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the national meeting of the National Council on Measurement in Education, New Orleans.
- Spray, J.A., & Reckase, M.D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized

test. *Journal of Educational and Behavioral Statistics*, 21, 405-414.

Van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63, 201-216.

Veerkamp, W.J.J. & Berger, M.P.F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, 22, 203-226.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.



## **Samenvatting**

Computergestuurde adaptieve toetsen (CATs) zijn geïndividualiseerde toetsen die in een geautomatiseerde omgeving worden afgenomen. Met een CAT wordt een vaardigheid van een persoon gemeten met een selectie opgaven uit een beschikbare itembank. Die selectie van de opgaven vindt plaats tijdens de afname van de toets en wel zodanig dat het niveau van de opgaven wordt afgestemd op het tot dan gedemonstreerde vaardigheidsniveau van de kandidaat. Voordelen van CATs boven traditionele (lineaire) (computergestuurde) toetsen zijn de vergroting van de meetefficiëntie en de toetsing van elke kandidaat op zijn eigen niveau.

CATs maken gebruik van een gekalibreerde itembank: dit een verzameling opgaven, waarvoor een itemresponsmodel geldt en waarin de parameterwaarden van de items zijn vastgelegd. Omdat in operationele CATs de itemparameters als bekend worden verondersteld, is het van groot belang dat het kalibratie-onderzoek van de items adequaat plaatsvindt en aan hoge eisen voldoet. In hoofdstuk 2, 3 en 4 wordt aandacht besteed aan de kalibratie van itemparameters in itemresponsmodellen voor dichotoom scorebare items. De aandacht is daarbij gericht op die modellen waarin de schatting van de itemparameters ook met behulp van de conditionele maximum aannemelijkheidsmethode (CML) kan plaatsvinden.

De voorschriften volgens welke de start, de wijze van itemselectie, de bepaling van de vaardigheid en de beëindiging van een toets plaatsvinden, worden vastgelegd in het algoritme van een CAT. Optimale algoritmen zijn afhankelijk van het doel van de toets en of van de toepassingscontext. In de hoofdstukken 5, 6 en 7 worden elementen van algoritmen behandeld passend bij specifieke toetsdoelen of praktische randvoorwaarden.

In itemresponsmodellen van het Rasch type wordt de CML methode voor het schatten van de itemparameters vanwege een aantal goede statistische eigenschappen veel gebruikt. In hoofdstuk 2 wordt het informatieverlies bij het toepassen van deze methode behandeld. Daarbij wordt gebruik gemaakt van het

begrip F-informatie, waarvan de geldigheid van enkele belangrijke eigenschappen wordt aangetoond. Dit begrip maakt het mogelijk voorwaarden te specificeren waaronder er geen verlies van informatie is in bepaalde klassen van modellen. Bovendien kan op basis van dit begrip een eventueel verlies aan informatie gekwantificeerd worden, waarvoor een gestandaardiseerde determinant van de F-informatiematrix wordt gebruikt. Voor het dichotome Raschmodel worden in dit hoofdstuk in detail de uitwerkingen gegeven voor het vergelijken van verschillende schattingsmethoden. Aangetoond wordt dat bij CML schatting van de itemparameters altijd enige informatie verloren wordt. Maar vergeleken met de alternatieve schattingsmethoden, gebaseerd op de gezamenlijke maximale aannemelijkheidsmethode (JML) en op de marginale maximale aannemelijkheidsmethode (MML), is het verlies klein. De gerapporteerde efficiency van de aanwezige informatie bij CML ten opzichte van bij JML en ook ten opzichte van bij MML is in de besproken voorbeelden altijd groter dan 92%. Bovendien neemt het gerapporteerde verlies af bij toenemende toetslengte: bij een toetslengte van meer dan 12 items is het minder dan 1%.

Doorgaans is het niet mogelijk om alle in een itembank op te nemen items bij alle proefpersonen af te nemen. Daarom worden voor de kalibratie de gegevens meestal verzameld in onvolledige test designs. Het gebruiken van de F-informatie voor de het vergelijken van de CML- en MML- methode voor het schatten itemparameters wordt in hoofdstuk 3 veralgemeend naar situaties met incomplete designs. Daarnaast wordt in dit hoofdstuk de relatie tussen de normalisatie van het Raschmodel en de op de F-informatie gebaseerde scalaire maat voor informatie en efficiënte uitgediept. Aangetoond wordt dat voor het vergelijken van schattingsmethoden de gebruikte informatie-efficiënte onafhankelijk is van de gekozen normalisatie van het model. Aan de hand van voorbeelden in enkele veel gebruikte onvolledige designtypen worden vergelijkingen uitgevoerd tussen CML- en MML schatten, maar ook vergelijkingen tussen het gebruiken van verschillende designs. De beschouwde designtypen zijn de itemketting ankerdesigns, de geblokte ketting ankerdesign en de gebalanceerde blokdesigns. Gevonden is dat zowel voor CML als voor

MML er altijd enige informatie wordt verloren in incomplete designs vergeleken met complete designs. Een algemene bevinding is dat met toenemende lengte van de gebruikte toetsboekjes de efficiëntie van een onvolledig ten opzichte van een volledige design, en ook de efficiëntie van CML vergeleken met MML toeneemt. Een verschil tussen CML en MML is er gevonden met betrekking tot het effect van de lengte van een toetsboekje: voor hele korte toetsboekjes is er een substantieel verlies (ongeveer 35 %) in informatie bij CML schatten, terwijl dit slechts 10% is bij MML schatten. Echter bij toenemende lengte van de toetsboekjes, vanaf ongeveer 10 items, verdwijnen de verschillen tussen CML en MML snel.

In de onvolledige designs die in hoofdstuk 3 zijn besproken vindt er een aselechte toewijzing plaats van de toetsboekjes aan de kandidaten. In hoofdstuk 4 wordt de rechtvaardiging van de kalibratie met CML en MML in meer algemene onvolledige designs behandeld. Het stochastisch karakter van de ontbrekende gegevens speelt daarbij een grote rol. Naast gerandomiseerde onvolledige designs worden meerfasen en groepsgerichte onvolledige designs besproken. Voor de rechtvaardiging van de MML toepassing wordt gebruik gemaakt van de algemene theorie van Rubin over de “ignorability” van het proces dat de ontbrekende gegevens veroorzaakt. Eerdere resultaten uit de literatuur worden in dit kader uitgewerkt en gerecapituleerd. De rechtvaardiging voor correcte toepassing van CML procedures wordt vastgesteld door het in CML genegeerde deel van de totale aannemelijkheid te beschouwen. In het hoofdstuk worden voorbeelden gegeven van onjuiste toepassing van CML en MML procedures in bepaalde designs. Zo wordt gedemonstreerd dat toepassing van MML in groepsgerichte designs, zonder simultane schatting van de volledige populatiestructuur, leidt tot foute itemparameterschattingen. Met CML worden onzuivere schattingen van de itemparameters verkregen met gegevens die verzameld zijn in meerfasen designs.

In de hoofdstuk 5, 6 en 7 behandelt dit proefschrift onderwerpen die direct toegepast (kunnen) worden bij de ontwikkeling en productie van computer-gestuurde adaptieve toetsen. Het betreft studies die elementen van het algoritme

van de CATs behandelen.

In hoofdstuk 5 wordt ingegaan op het vraagstuk welke algoritmen gebruikt kunnen worden in een CAT waarbij het toetsdoel het classificeren van kandidaten in één van drie te onderscheiden categorieën is. In CATs wordt voor de vaststelling van de vaardigheid doorgaans gebruikt gemaakt van statistische schatting. In deze dissertatie is de schatting altijd gebaseerd op een maximale gewogen aannemelijkheidsmethode. Als alternatief kan bij het gegeven toetsdoel ook een statistische toetsingsmethode, de sequential probability ratio test (SPRT), worden gebruikt. In dit hoofdstuk worden deze twee methoden voor de bepaling van de vaardigheid gecombineerd met itemselectiemethoden die gebaseerd zijn op maximale (Fisher) informatie (MI). Op basis van simulatiestudies, waarbij gebruik gemaakt wordt van een gekalibreerde itembank met 250 wiskunde-opgaven, wordt de meetkwaliteit van de voorgestelde algoritmen geëvalueerd. Uit deze studies blijkt dat in elk voorgesteld CAT algoritme een reductie haalbaar is van op zijn minst 22% van het gemiddeld aantal benodigde opgaven om een zelfde nauwkeurigheid te bereiken als bij een bestaande papieren twee-fasen plaatsingstoets. Uit de vergelijkingen blijkt dat algoritmen gebaseerd op toepassing van de SPRT een veelbelovend alternatief zijn voor algoritmen gebaseerd op statistische schatting. Naast een kleine afname in het gemiddeld aantal benodigde items, vraagt de methode gebaseerd op de SPRT veel minder rekentijd dan de schattingsmethode tijdens de toetsafname. Tenslotte is geconstateerd dat het opleggen van extra randvoorwaarden op de MI itemselectie in de vorm van inhoudscontrole en afnamecontrole nauwelijks negatieve invloed heeft op de kwaliteit van de algoritmen.

In hoofdstuk 6 worden CATs die de SPRT in het algoritme gebruiken nader bestudeerd. Bij eerdere toepassing hiervan in CATs zijn altijd itemselectiemethoden gebruikt die ofwel aselekt items uit de itembank trekken dan wel gebaseerd zijn op het maximaliseren van de (Fisher) informatie. Bij het selecteren van items met maximale Fisher informatie bij de lopende vaardigheidsschatting wordt de nauwkeurigheid van de vaardigheid van de kandidaat geoptimaliseerd. In dit hoofdstuk wordt een itemselectiemethode

gepresenteerd die theoretisch beter aansluit bij de toepassing van een statistisch toetsingsprobleem en die daarmee een algoritme voor classificatiebeslissingen gebaseerd op de SPRT zou kunnen verbeteren. De itemselectie is gebaseerd op het maximaliseren van de Kullback-Leibler informatie, waarvoor de uitdrukkingen voor het 1-, het 2- en het 3-paramater logistisch testmodel worden gegeven. In simulatiestudies met dezelfde wiskunde itembank als die in hoofdstuk 5 werd gebruikt zijn een aantal varianten van selectiemethoden gebaseerd op de Fisher en op de Kullback -Leibner informatie vergeleken. Hierbij werden zowel classificatieproblemen in twee categorieën als in drie categorieën beschouwd. Het algemene resultaat van deze studies is dat adaptieve algoritmen die de SPRT gebruiken in combinatie met itemselectie gebaseerd op de Kullback-Leibner informatie soms beter maar nooit slechter presteren dan de combinatie met selectie op basis van de Fisher informatie. De gerapporteerde verschillen zijn evenwel klein.

In het laatste hoofdstuk wordt onderzocht in hoeverre het mogelijk is de moeilijkheidsgraad van de toetsing in CATs aan te passen aan in de toetspraktijk levende wensen. CATs worden voor elk individu, onder een beperkt aantal praktische condities, vanuit meetoogpunt optimaal samengesteld. Als gebruik gemaakt wordt van itembanken die gekalibreerd zijn met het één- (1pl) of het twee-parameter logistisch testmodel (2pl), dan heeft dat als consequentie dat elke persoon ongeveer 50% van de items juist beantwoord. Bij het toepassen van CATs, bij bijvoorbeeld kleuters, zou deze moeilijkheid ongewenste bijeffecten kunnen hebben. In hoofdstuk 7 worden twee methoden voor itemselectie besproken en geëvalueerd, die het mogelijk maken eventueel gemakkelijker of moeilijker CATs te maken. Een uit de literatuur bekende methode, gebaseerd op het selecteren van items waarvan het gewenste kanspunt het dichtst bij de vaardigheidsschatting ligt, blijkt bij 1pl itembanken goed te functioneren, maar niet bij een 2pl itembank. Een alternatieve methode wordt gegeven. Deze methode selecteert items met maximale informatie bij een vaardigheid die ten opzichte van de lopende vaardigheidsschatting verschoven is. Deze methode blijkt voor beide itemresponsmodellen goede resultaten te geven. Met deze



methode kan bij itembanken met een omvang die veel in de praktijk voorkomt, zonder groot verlies aan nauwkeurigheid, de gewenste moeilijkheid van een CAT gekozen worden zodanig dat kandidaten tussen 30% en 70% van de items goed maken.