Measurement and Research Department Reports 92-3

Item Calibration in Incomplete Testing Designs

T.J.H.M Eggen
N.D. Verhelst

Cito

Measurement and Research Department Reports 92-3

Item Calibration in Incomplete Testing Designs

T.J.H.M. Eggen
N.D. Verhelst

Cito
Arnhem, 1992

Abstract

The paper 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. It is 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. In CML this is not possible and the justification can be established from the S-ancillarity condition of the neglected part of the likelihood. Incorrect uses of standard MML- and CML-algorithms are discussed.

Keywords: Random incomplete design, multistage testing design, targeted testing design, IRT, CML, MML, ignorability

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, efficiency often 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 (Mislevy & Bock, 1982), uses the marginal maximum likelihood (MML) approach in the one-, two- and three-parameter logistic testmodel and OPLM (Verhelst, 1992), 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 paper 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.

The paper addresses a topic which is related to a paper by Glas (1988) on item calibration and multistage testing in the Rasch model, to an ETS-report by Mislevy and Wu (1988) on estimating ability of students in experiments with missing data, and an article of Mislevy and Sheenan (1989) on the use of collateral information about students in item parameter estimation.

Item Response Theory (IRT)

In IRT we consider the random vector, the response pattern $\boldsymbol{X} = (X_{ij})$, $i=1,...,n; j=1,...,k$, 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)

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

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

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

IRT models further assume local independence between item responses. If we denote $\boldsymbol{X}_i = (X_{ij}), j=1,...,k$ as the response vector of student i and Y_i as any other measured characteristic of student i (possibly multivariate but not functionally dependent on $\boldsymbol{X}_i$):

$$P_{\theta_i,\boldsymbol{\beta}}(\boldsymbol{x}_i \mid y_i) = P_{\theta_i,\boldsymbol{\beta}}(\boldsymbol{x}_i) = \prod_j P_{\theta_i,\beta_j}(x_{ij}), \quad (i=1,...,n), \tag{2}$$

in which $\boldsymbol{\beta} = (\beta_j), j=1,...,k$. Local independence means that item responses are independent given the ability and furthermore that given the ability the responses are independent of all other person characteristics. Finally independence of item responses between students is assumed, which can be established for example, via random sampling from a population, which means (with $\boldsymbol{\theta} = (\theta_i), i=1,...,n$),

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

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

$$L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{x}) = P_{\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{x}) \tag{4}$$

is maximized with respect to the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. It is well known that because of the incidental parameters θ_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.

Conditional Maximum Likelihood (CML)

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

$$P_{\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{x}) = \prod_i P_{\boldsymbol{\beta}}(\boldsymbol{x}_i \mid s(\boldsymbol{x}_i)) \cdot P_{\theta_i,\boldsymbol{\beta}}(s(\boldsymbol{x}_i)) \qquad \text{if } S(\boldsymbol{X}) = s(\boldsymbol{x}) \ , \tag{5}$$

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

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

In CML estimation of the item parameters only random variations of the observations, fixing (given) the values of the conditioning statistics $s(\boldsymbol{x}_i)$, are considered. The justification of this depends on whether all random variation that is relevant to the problem (here estimating the item parameters $\boldsymbol{\beta}$) 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 $S(\boldsymbol{X}_i)$ would be completely independent of the item parameters $\boldsymbol{\beta}$, the justification would be obvious. In that case the distribution of $S(\boldsymbol{X}_i)$ is said to be ancillary with respect to $\boldsymbol{\beta}$. However this condition is not fulfilled in our situation. But Andersen (1973) has shown that it is sufficient for the distribution of the sufficient statistic $S(\boldsymbol{X}_i)$ to be (weakly) or S-ancillary with respect to $\boldsymbol{\beta}$. In the latter the resulting CML estimators of $\boldsymbol{\beta}$ are, under mild regularity conditions, consistent, and asymptotically normally distributed

and efficient. S-ancillarity can be interpreted as that the distribution of the sufficient statistic $S(\boldsymbol{X}_i)$ does not contain any information on the item parameters that is not completely dependent on the specification of the ability parameter. In other words, we cannot learn anything in the data about the item parameters from the sole observation of the sufficient statistic.

It should be noted that while S-ancillarity of the sufficient statistic has shown to be the key condition for the consistency and asymptotic normality and efficiency of CML estimators, this does not imply that S-ancillarity is a necessary condition in order to obtain CML estimators. Conditioning on a statistic not having this property, could well result in CML estimators, which are in some sense satisfactory, without having all the featured properties. With possible lost of efficiency, conditioning could well be the only simple way to obtain satisfactory estimators in some problems. However, in this paper the justifiability of CML estimation of item parameters is restricted to mean that the before mentioned properties of the estimators yield. Which is the case when the S-ancillarity condition is fulfilled. In a forthcoming paper by Eggen a more exact recast of the justifiabilty of CML estimation is presented.

Although intuitively S-ancillarity may be an appealing concept, it is not easy to show in general that a distribution has this property. However, because we consider for CML only models belonging to the exponential family, necessary and sufficient conditions for S-ancillarity are easily checked (Andersen, 1973). It is well known that the Rasch model is an exponential family model with minimal sufficient statistic $S(\boldsymbol{X}_i) = \sum_j \boldsymbol{X}_{ij}$, the sum score of a student i on the items in the test, whose distribution is S-ancillary with respect to $\boldsymbol{\beta}$. Defining $\delta(s,\boldsymbol{\beta})$, the elementary symmetric functions, by

$$\delta(s_i,\boldsymbol{\beta}) = \sum_{\{x_i \mid \sum_j x_{ij}=s_i\}} \exp(-\sum_j \beta_j x_{ij}), \tag{7}$$

the condition to meet for S-ancillarity in this model boils down to

$$\frac{\partial \log \delta(s,\boldsymbol{\beta})}{\partial \beta_j} = a(\boldsymbol{\beta}).s + b(\boldsymbol{\beta}), \quad (j=1,\ldots,k), \tag{8}$$

with $a(\boldsymbol{\beta})$ and $b(\boldsymbol{\beta})$ being only functions of $\boldsymbol{\beta}$.

This condition is easily checked to be fulfilled in the Rasch model where CML estimation is common good practice. Verhelst and Eggen (1989) applied the CML approach to the general class of one-parameter logistic models. Here $S_2(\boldsymbol{X}_i)=\sum_j a_j X_{ij}$, the weighted score on a test, is minimally sufficient for θ_i and CML estimation of the item parameters $\boldsymbol{\beta}$ is possible. In these models $a_j, (j=1,...,k)$ is a fixed discrimination-index of item j.

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 of 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.

Marginal Maximum Likelihood (MML)

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

$$P_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{x}) = \int_{\boldsymbol{\theta}} P_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{x},\boldsymbol{\theta})\, d\boldsymbol{\theta} = \prod_i \int_{\theta_i} P_{\boldsymbol{\beta}}(\boldsymbol{x}_i \mid \theta_i)\, g_{\boldsymbol{\gamma}}(\theta_i) d\theta_i \,, \tag{9}$$

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

In MML estimation the item parameters $\boldsymbol{\beta}$ are simultaneously estimated with the parameter $\boldsymbol{\gamma}$ of the ability distribution by maximizing the marginal probability of the observed response pattern $\boldsymbol{x}$ (the marginal likelihood function) with respect to the parameters, that is,

$$L_m(\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{x}) = \prod_i \int_{\theta_i} P_{\boldsymbol{\beta}}(\boldsymbol{x}_i \mid \theta_i) \cdot g_{\boldsymbol{\gamma}}(\theta_i) d\theta_i \,. \tag{10}$$

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 θ is normal with $\gamma = (\mu, \sigma^2)$. Bock and Aitkin (1981) were the first to give computational procedures for maximizing (10) using the EM-algorithm.

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 $\boldsymbol{U} = (U_1,...,U_m)$ a vector random variable with probability density function $f_\tau(\boldsymbol{u})$. τ is a vector parameter, on which we want to draw inferences on the basis of the data, a sample realization $\boldsymbol{u}$. 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, $\boldsymbol{M} = (M_1,...,M_m)$ is defined, indicating whether a variable U_j, is actually observed, $m_j = 1$, or not observed, $m_j = 0$. The observed value of $\boldsymbol{M}$ $(\boldsymbol{m})$ effects a partition of the vector random variable $\boldsymbol{U}$ and its observed value, corresponding with $m_j = 0$ for missing and $m_j = 1$ for observed data: $\boldsymbol{U} = (\boldsymbol{U}_{obs}, \boldsymbol{U}_{mis})$ and $\boldsymbol{u} = (\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis})$. Respectively the sets of indices of observed and not observed variables are:

$$obs = \{j \mid m_j=1\} \quad \text{and} \quad mis = \{j \mid m_j=0\} . \tag{11}$$

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

$$h_\phi(\boldsymbol{m} \mid \boldsymbol{u}) = P_\phi(\boldsymbol{M} = \boldsymbol{m} \mid \boldsymbol{U} = \boldsymbol{u}) , \tag{12}$$

which is defined as the distribution corresponding to the process that causes the missing data, with ϕ a possibly vector valued parameter. In general ϕ will be dependent on the parameter of interest τ.

In the presence of missing data we have a sample realization of $\boldsymbol{M}$ and $\boldsymbol{U}_{obs}$ and the basis for inference should be their joint distribution:

$$\int_{u_{mis}} f_{\tau,\phi}(u,m)\, du_{mis} = \int_{u_{mis}} f_{\tau}(u) \cdot h_{\phi}(m \mid u)\, du_{mis} \cdot \qquad (13)$$

However, because we are only interested to infer on the parameter τ of the distribution of the partially observed U, a possible approach could be to ignore the process that causes the missing data in the inference. Following Rubin (1976), ignoring the process that causes missing data means: (a) fixing the random variable M at the observed pattern of missing data m and (b) assuming that the values of the observed data u_{obs} are realizations of the marginal density of U_{obs} :

$$\int_{u_{mis}} f_{\tau}(u)\, du_{mis} \cdot \qquad (14)$$

So, when we ignore the process that causes the missing data, not all possible random variation in the data due to, sampling of M and U_{obs} , is considered, but only random variation due to U_{obs} fixing the random variable M at the particularly observed pattern m. The generally more convenient form (14) is used in stead of (13) in the inference on τ.

It will be clear that ignoring the missing data process does not necessarily lead to a correct inference on τ. Firstly, we possibly disregard the dependence of ϕ on τ. Secondly, it is understood that the data u_{obs} are in fact realizations of the conditional density of U_{obs} given the random variable M took the fixed value m :

$$\int_{u_{mis}} f_{\lambda}(u \mid m)\, du_{mis}, \qquad (15)$$

with λ the possibly vector valued parameter of this distribution, which could depend on τ (the parameter of the distribution of U) as well as on the parameter of the missing data process ϕ .

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 τ. We will only consider the sufficient conditions in likelihood inference, because these suffice for our arguments later. These conditions are conditions on the distribution $h_{\phi}(m \mid u)$ corresponding to the process that causes missing data. Defined are:

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

$$h_{\phi}(m \mid u_{obs}, u_{mis}) = h_{\phi}(m \mid u_{obs}) \text{ for all } u_{mis}, \tag{16}$$

that is, the missingness of the data does not depend on the missing values of U_{mis}, but may depend on the observed values of U_{obs}.

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

$$h_{\phi}(m \mid u_{obs}, u_{mis}) = h_{\phi}(m) \text{ for all } u_{mis} \text{ and } u_{obs}. \tag{17}$$

Note that MCAR implies MAR.

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

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 τ.
This means that in stead of using the full-likelihood

$$L(\tau, \phi; u_{obs}, m) = f_{\tau,\phi}(u_{obs}, m), \tag{18}$$

the simple likelihood function

$$L(\tau; u_{obs}) = f_{\tau}(u_{obs}) = \int_{u_{mis}} f_{\tau}(u)\, du_{mis}, \tag{19}$$

can be used. Under MAR and D this is easily seen to be true since the joint distribution of the U_{obs} and M equals in this case:

$$
\begin{aligned}
f_{\tau,\phi}(u_{obs},m) &= \int_{u_{mis}} f_{\tau}(u_{obs},u_{mis}) \,.\, h_{\phi}(m \mid u_{obs},u_{mis})\, du_{mis} \\
&= h_{\phi}(m \mid u_{obs}) \int_{U_{mis}} f_{\tau}(u_{obs},u_{mis})\, du_{mis} \\
&= h_{\phi}(m \mid u_{obs}) \,.\, f_{\tau}(u_{obs}).
\end{aligned}
\tag{20}
$$

In (20) the second equality holds because of MAR (16) and the distribution of the data factors in two terms. When D holds these factors can be maximized separately. So in inferring on τ it is equivalent to use the simple likelihood function (19). Ignoring the process that causes missing data is of course also justified if the stronger condition MCAR, in stead of MAR, (and D) is met.

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 $t=1,...,T$. From the total item pool of k items, subsets of $k_t, (t=1,...,T)$ 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 this is not strictly necessary (see Glas, 1989). 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 $\boldsymbol{R}_i$ can take T values:

$$
\boldsymbol{r}_t = perm_t\,(\mathbf{1}_{k_t}, \mathbf{0}_{k-k_t}), \qquad (t=1,...,T)\,. \tag{21}
$$

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. Which is in accordance with the general design random variable $\boldsymbol{M}_i$ defined in the preceding section. The item indicator variable $\boldsymbol{R}$ of the total sample of students (size n) consists of the vectors $\boldsymbol{R}_i$ of each student in the sample: $\boldsymbol{R} = (\boldsymbol{R}_1,...,\boldsymbol{R}_n)$.

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 be 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 ϕ_t such that $\sum_{t=1}^{T} \phi_t = 1$. The distribution of the item indicator variable $\boldsymbol{R}_i$ is given by:

$$P(\boldsymbol{R}_i = \boldsymbol{r}_{t_i}) = \phi_{t_i} \text{ with } t_i \in \{1,...,T\},\ (i=1,...,n). \tag{22}$$

A priori fixed incomplete designs are of course special cases of random incomplete designs ($P(\boldsymbol{R}_i=\boldsymbol{r}_t) = 1$ or 0).

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. In general, all students in the sample take the first stage test which is of medium difficulty. Students with high scores on this first stage 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. An example is given in Figure 1.

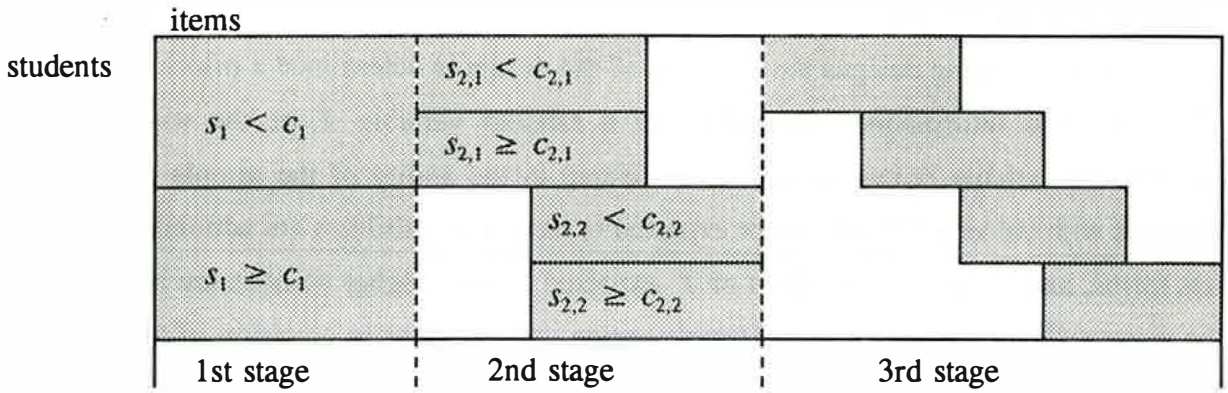

Figure 1: Example of a multistage testing design.

In Figure 1, $s.$ indicates the scores of students on a set of items which are in each stage 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.

In a multistage testing design, as in a random incomplete design, the item indicator variable for every student $\boldsymbol{R}_i$ can take as many values as there are tests T (see 21). The distribution of $\boldsymbol{R}_i$ has always the following form:

$$\left.\begin{array}{l} P(\boldsymbol{R}_i = \boldsymbol{r}_t \mid \boldsymbol{x}_{obs,i} \in C_t) = 1 \\ P(\boldsymbol{R}_i = \boldsymbol{r}_t \mid \boldsymbol{x}_{obs,i} \notin C_t) = 0 \end{array}\right\} , (t=1,...,T) . \tag{23}$$

If a function of observed item scores $\boldsymbol{x}_{obs,i}$ meets a criterion for getting test t, the item indicator variable $\boldsymbol{R}_i$ takes the value $\boldsymbol{r}_t$ with probability 1. If the criterion is not met the probability is 0.

In the example the function is the sumscore on a subset of the items and the criterion is given by the cut-offs:

$$C_1 = \{(\boldsymbol{x}_{obs}) \mid s_1(\boldsymbol{x}_{obs}) < c_1 , s_{2,1}(\boldsymbol{x}_{obs}) < c_{2,1}\} . \tag{24}$$

The criterion is met for all those response vectors, for which the score on the first stage test is smaller than c_1, and the score on the second stage test is smaller than $c_{2,1}$.

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 $\boldsymbol{Y}$ of the students. This background variable is usually positively related to the ability of the sample of students. Students with values of $\boldsymbol{Y}$ which are expected to have lower abilities are administered easier test forms, and students with values of $\boldsymbol{Y}$ 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 $\boldsymbol{Y}$ of the students is categorical (or categorized), taking (or distinguishing) T values: $y_1,\ldots,y_T$. In targeted testing, for each value of $\boldsymbol{Y}$ a different subset from the total item pool is administered to the students. The variable $\boldsymbol{Y}$ 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 $\boldsymbol{Y}$ is only used in the assignment of items or tests to students and not in the sampling of students. Second, the $\boldsymbol{Y}$ 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 $\boldsymbol{Y}$ 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 $\boldsymbol{Y}$ itself. There is, for instance, no interest to have estimates of the parameters of the ability distribution for each distinguished level of $\boldsymbol{Y}$. Here the students are sampled from one population with no regard to the values of $\boldsymbol{Y}$. 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 $\boldsymbol{Y}$ itself. A situation often occurring is that $\boldsymbol{Y}$ 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

only as separate samples from more than one population. The variable $\boldsymbol{Y}$ is used to stratify the total population in the 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). Note that these two sampling roles of $\boldsymbol{Y}$ can also be distinguished in complete testing designs: complete testing with a student sample from one population (CTOP) and complete testing with student samples from multiple (sub)populations (CTMP).

In targeted designs the item indicator variable $\boldsymbol{R}_i$ for every student can again take as many values as there are tests (see 21). The distribution of $\boldsymbol{R}_i$ is given by:

$$P(\boldsymbol{R}_i=\boldsymbol{r}_t) = P(\boldsymbol{Y}_i=\boldsymbol{y}_t) \ , \quad (t=1,...,T) \ , \tag{25}$$

or similar to (23) by

$$\left.\begin{aligned} P(\boldsymbol{R}_i = \boldsymbol{r}_t \mid \boldsymbol{Y}_i = \boldsymbol{y}_t) &= 1 \\ P(\boldsymbol{R}_i = \boldsymbol{r}_t \mid \boldsymbol{Y}_i \neq \boldsymbol{y}_t) &= 0 \end{aligned}\right\}, \ (t=1,...,T). \tag{26}$$

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 $\boldsymbol{R}$ 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 preceding section. We will explore the applicability of the general framework of Rubin (1976) for 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 $\boldsymbol{U}_{obs,i}$ and $\boldsymbol{U}_{mis,i}$, $i = 1,...,n$, $\boldsymbol{U} = (\boldsymbol{U}_{obs},\boldsymbol{U}_{mis})$, with

$\boldsymbol{U}_{obs} = (\boldsymbol{U}_{obs,1},...\boldsymbol{U}_{obs,n})$ and $\boldsymbol{U}_{mis} = (\boldsymbol{U}_{mis,1},...\boldsymbol{U}_{mis,n})$. The missing data indicator is $\boldsymbol{M} = (\boldsymbol{M}_1,...,\boldsymbol{M}_n)$, in which every element $\boldsymbol{M}_i$ is a vector of the same length as there are variables (observed and unobserved).

The Marginal Model and Missing Data

The MML estimation procedure for complete data, see (9) and (10), 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. Note that the variable on which we want to infer is $\boldsymbol{U} = (\boldsymbol{X},\boldsymbol{\theta}) = (\boldsymbol{X}_1,\theta_1,...,\boldsymbol{X}_n,\theta_n)$ in which $\boldsymbol{X}_i$ 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 $\boldsymbol{X}_i$ are always observed and the θ_i are always missing. So we have for every student i a degenerated design distribution, that equals its item indicator distribution

$$P(\boldsymbol{M}_i = (\boldsymbol{1}_k,0)) = P(\boldsymbol{R}_i = (\boldsymbol{1}_k)) = 1 \ , \quad (i=1,...,n) \ . \tag{27}$$

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

$$\boldsymbol{U}_{obs,i} = \boldsymbol{X}_i \quad \text{and} \quad \boldsymbol{U}_{mis,i} = \theta_i \ , \quad (i=1,...,n) \ . \tag{28}$$

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

$$\int_{U_{mis}} f_{\tau}(u)\, du_{mis} = \int_{\theta} P_{\beta,\gamma}(x,\theta)\, d\theta = \prod_i \int_{\theta_i} P_{\beta}(x_i \mid \theta_i)\, g_{\gamma}(\theta_i)\, d\theta_i \ . \tag{29}$$

Which is identical to (10). The justification of using MML for estimation the parameters $(\boldsymbol{\beta},\boldsymbol{\gamma})$ can thus also be deduced from the general framework of Rubin for inference in the presence of missing data.

MML in Random Incomplete Designs

Again we have here

$$\boldsymbol{U} = (\boldsymbol{X},\theta) \text{ and } \boldsymbol{\tau} = (\boldsymbol{\beta},\boldsymbol{\gamma}) . \tag{30}$$

The design distribution follows from (22):

$$P(\boldsymbol{M}_i = (r_{t_i}, 0)) = P(\boldsymbol{R}_i = r_{t_i}) = \phi_{t_i} , \quad (i=1,...,n) . \tag{31}$$

The partition established by an observation $\boldsymbol{m}_i$ of the design variable is given by:

$$\left.\begin{aligned} \boldsymbol{U}_{obs,i} &= \boldsymbol{X}_{obs,i} \\ \boldsymbol{U}_{mis,i} &= (\boldsymbol{X}_{mis,i}, \theta_i) \end{aligned}\right\} , \quad (i=1,...,n). \tag{32}$$

Because the design distribution does not depend on any of the missing nor the observed responses, it meets the conditions of MCAR and D. The likelihood inference can be based on the marginal distribution of the observations:

$$\begin{aligned} &\prod_i \int_{x_{mis,i}} \int_{\theta_i} P_{\boldsymbol{\beta}}(\boldsymbol{x}_{obs,i}, \boldsymbol{x}_{mis,i} \mid \theta_i) \cdot g_{\boldsymbol{\gamma}}(\theta_i)\, d\theta_i\, d\boldsymbol{x}_{mis,i} = \\ &\prod_i \int_{\theta_i} P_{\boldsymbol{\beta}}(\boldsymbol{x}_{obs,i} \mid \theta_i) \cdot g_{\boldsymbol{\gamma}}(\theta_i)\, d\theta_i . \end{aligned} \tag{33}$$

The ignorability principle can be applied.

Note that if we indicate by $\boldsymbol{n}_t$ the number of students taking test t with $\sum_{t=1}^{T} \boldsymbol{n}_t = \boldsymbol{n}$ and define $\boldsymbol{\beta}_{(t)}$ as the k_t-vector of the item parameters of the items in test t, we can rewrite (33) as

$$\prod_i \int_{\theta_i} P_{\boldsymbol{\beta}}(\boldsymbol{x}_{obs,i} \mid \theta_i) \cdot g_{\boldsymbol{\gamma}}(\theta_i)\, d\theta_i = \prod_{t=1}^{T} \prod_{i=1}^{n_t} \int_{\theta_i} P_{\boldsymbol{\beta}_{(t)}}(\boldsymbol{x}_{obs,i} \mid \theta_i) \cdot g_{\boldsymbol{\gamma}}(\theta_i)\, d\theta_i . \tag{34}$$

The marginal likelihood in the incomplete data case is written as a product of T complete data marginal likelihoods with overlapping items.

MML in Multistage Testing Designs

U, $U_{obs,i}$, $U_{mis,i}$ and τ are as in (30) and (32). The distribution of M_i follows from (23):

$$P(M_i = (r_{t_i}, 0) \mid x_{obs,i}) = P(R_i = r_{t_i} \mid x_{obs,i}) = 1 \quad \text{or } 0 \ , \ (i=1,\ldots,n). \tag{35}$$

Because the design distribution (35) only depends on observed responses the MAR condition is fulfilled. Condition D is also clearly met. Therefore ignorability holds in multistage testing designs and MML can be applied using the marginal distribution of the observations. This can readily be checked by considering the distribution of (U_{obs}, M) needed for the full likelihood:

$$\begin{aligned}
&\int_{u_{mis}} P_{\tau,\phi}(u, m)\, du_{mis} = \int_{\theta} \int_{x_{mis}} P_{\beta,\gamma,\phi}(x_{obs},x_{mis},\theta, m)\, dx_{mis}\, d\theta = \\
&\int_{\theta} \int_{x_{mis}} P_{\beta,\gamma}(x_{obs},x_{mis},\theta) \cdot h_{\phi}(m \mid x_{obs},x_{mis},\theta)\, dx_{mis}\, d\theta = \\
&h_{\phi}(m \mid x_{obs}) \int_{\theta} P_{\beta,\gamma}(x_{obs},\theta)\, d\theta = \\
&\prod_i h_{\phi}(m_i \mid X_{obs,i}) \prod_i \int_{\theta_i} P_{\beta}(X_{obs,i} \mid \theta_i) \cdot g_{\gamma}(\theta_i)\, d\theta_i \ .
\end{aligned} \tag{36}$$

In (36) the third equality holds because of MAR resulting in a factorization of the full likelihood in a term independent of (β,γ) and the marginal distribution of X_{obs}. So just considering the marginal distribution of X_{obs} will thus give the correct maximum likelihood estimates of β and γ.

MML in Targeted Testing Designs

Mislevy and Sheenan (1989) present a more general discussion of the effect of using or not using (ignoring) the background information of the students in item calibration, depending on the role this background variable $\boldsymbol{Y}$ has in the testing design. We will reconsider their results. They assume, as we did before, $\boldsymbol{Y}$ to be a categorical (or categorized) variable taking one of $\boldsymbol{L}$ values, establishing a division of the total student population in $\boldsymbol{L}$ subpopulations. The value of $\boldsymbol{Y}$ for student i is defined as $\boldsymbol{y}_i = (y_{i1},...,y_{iL})$ with $y_{i\ell} = 1$ if student i is associated with subpopulation ℓ and 0 if not, $\ell = 1,...,L$. If $y_{i\ell} = 1$ we will alternatively write $\boldsymbol{y}_i = y^{(\ell)}$. The ability distribution $g_{\boldsymbol{Y}}(\theta)$ of the total population in that case can be expressed as a finite mixture of $\boldsymbol{L}$ subpopulation ability distributions:

$$g_{\boldsymbol{Y}}(\theta) = \sum_{\ell=1}^{L} P(\theta, \boldsymbol{Y} = y^{(\ell)}) = \sum_{\ell=1}^{L} P(\theta \mid \boldsymbol{Y} = y^{(\ell)}) \cdot P(\boldsymbol{Y} = y^{(\ell)})$$
$$= \sum_{\ell=1}^{L} g_{\boldsymbol{\gamma}_\ell}(\theta) \cdot \pi_\ell \,, \tag{37}$$

in which $\boldsymbol{\gamma}_\ell$ is the possibly vector valued parameter of the ability distribution in subpopulation ℓ and π_ℓ the proportion of subpopulation ℓ in the total population.

Before discussing targeted testing designs we will first consider using or ignoring the background information of the students in complete testing for the two roles of $\boldsymbol{Y}$ in sampling the students as described before: CTOP and CTMP.

In CTOP we have a random sample from the total population, with ability distribution $g_{\boldsymbol{Y}}(\theta)$, and besides the item response pattern $\boldsymbol{X}$ the values of $\boldsymbol{Y}$ are observed. Using the background information in the item calibration with marginal maximum likelihood methods, makes it possible to estimate simultaneously the item parameters $\boldsymbol{\beta}$, with the parameters of the ability distributions of the distinguished subpopulations. This generalization of the standard application of MML follows readily. The simultaneous probability of the response vector $\boldsymbol{X}_i$ and the background variable $\boldsymbol{Y}_i$ of a randomly sampled student i is given by:

$$P_{\beta,\gamma}(x_i, Y_i=y^{(\ell)}) = \int_{\theta_i} P_{\beta,\gamma}(x_i, Y_i=y^{(\ell)},\theta_i)\, d\theta_i =$$

$$\int_{\theta_i} P_{\beta}(x_i \mid Y_i=y^{(\ell)},\theta_i)\,.\, P_{\gamma}(\theta_i \mid Y_i=y^{(\ell)})\,.\, P(Y_i=y^{(\ell)})\, d\theta_i =$$

$$\int_{\theta_i} P_{\beta}(x_i \mid \theta_i)\,.\, g_{\gamma_\ell}(\theta_i).\pi_\ell\, d\theta_i = \qquad (38)$$

$$\prod_{\ell=1}^{L} \int_{\theta_i} \left[P_{\beta}(x_i \mid \theta_i)\,.\, g_{\gamma_\ell}(\theta_i).\pi_\ell\right]^{y_{i\ell}} d\theta_i =$$

$$\prod_{\ell=1}^{L} \pi_\ell^{y_{i\ell}} \,.\, \prod_{\ell=1}^{L} \int_{\theta_i} \left[P_{\beta}(x_i \mid \theta_i)\,.\, g_{\gamma_\ell}(\theta_i)\right]^{y_{i\ell}} d\theta_i \,.$$

The likelihood of the total sample is given by:

$$L(\beta,\gamma,\pi;x,y) = \prod_{i=1}^{n} P_{\beta,\gamma}(x_i, Y_i = y^{(\ell)}) =$$

$$\prod_{i=1}^{n}\prod_{\ell=1}^{L} \pi_\ell^{y_{i\ell}} \,.\, \prod_{i=1}^{n}\prod_{\ell=1}^{L} \int_{\theta_i} \left[P_{\beta}\,(x_i \mid \theta_i)\,.\, g_{\gamma_\ell}(\theta_i)\right]^{y_{i\ell}} d\theta_i \,. \qquad (39)$$

From (39) it is seen that the likelihood function consist of a term only dependent on the proportions π_ℓ of the subpopulations in the total population and a term which is a product of L ordinary marginal likelihood functions because there is always exact one ℓ for which $y_{i\ell}=1$. Standard maximum likelihood estimates $\hat{\pi}_\ell, \ell=1,...,L$ of the proportions can be obtained from the first part. Maximizing the second term with respect to $\gamma_\ell, \ell=1,...L$ and β will give estimates of L population parameters and the item parameters. Calibration using the background information in the CTOP case is thus a generalization of standard MML.

As in the standard MML this can also be described in Rubins framework. Here we have again for each student i a degenerated design distribution which equals its item indicator distribution:

$$P(M_i = (1_k,1,0)) = P(R_i = (1_k)) = 1\ ,\quad (i=1,...,n)\ . \qquad (40)$$

Compared to (27) the design vector variable M_i has one element more, indicating the observation of Y_i, which has always the value 1 in the CTOP case. The $(k+1)^{th}$ element indicates Y_i, the $(k+2)^{th}$ θ_i. The partition which the observed design variable m_i effects is

$$U_{obs,i} = (X_i, Y_i) \quad \text{and} \quad U_{mis,i} = \theta_i \ , \quad (i=1,...,n) \ . \tag{41}$$

Of course MCAR and D are met and marginalizing over the missing θ_i is justified according to the ignorability principle. If we do not use Y in this situation this means that (40) and (41) change into

$$P(M_i = (1_k,0,0)) = P(R_i = (1_k)) = 1 \ , \quad (i=1,...,n) \tag{42}$$

and

$$U_{obs,i} = X_i \quad \text{and} \quad U_{mis,i} = (\theta_i, Y_i) \ , \quad (i=1,...,n) \ . \tag{43}$$

The element of the design variable, indicating the observation of Y_i is now always 0 and Y is considered as a part of the missing data. Also in this situation the conditions for ignorability (MCAR and D) are met. So, in stead of using the full likelihood the marginal likelihood of the observed item responses can be used. For a randomly sampled student i we have:

$$\begin{aligned} P_{\beta,\gamma}(x_i) &= \int_{\theta_i} \sum_{l=1}^{L} P_{\beta,\gamma}(x_i, Y_i=y^{(l)},\theta_i) \, d\theta_i = \\ &\int_{\theta_i} P_\beta(x_i \mid \theta_i).\sum_{l=1}^{L} (g_{\gamma_l}(\theta_i).\pi_l) \, d\theta_i = \\ &\int_{\theta_i} P_\beta(x_i \mid \theta_i).g_\gamma(\theta_i) \, d\theta_i \ . \end{aligned} \tag{44}$$

For the total sample we have

$$P_{\beta,\gamma}(x) = \prod_{i=1}^{n} \int_{\theta_i} P_\beta(x_i \mid \theta_i) \cdot g_\gamma(\theta_i) \, d\theta_i \ , \tag{45}$$

which has the same form as the standard marginal likelihood (see 10). The parameters to be estimated are $\gamma=(\gamma_1,..,\gamma_\ell,..,\gamma_L,\pi_1,..,\pi_\ell,..,\pi_L)$ and $\beta=(\beta_1,...,\beta_k)$. So we can conclude that using and not using the background information in MML can be justified by Rubins principles in the CTOP case.

In the CTMP situation the background variable is used as a stratification variable: from every subpopulation ℓ, $\ell=1,...,L$, we have a random sample from $g_{\gamma_\ell}(\theta)$ with n_ℓ the number of observations in subpopulation ℓ 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 π_ℓ. These population proportions π_ℓ are not estimatable in this case but they must be known in advance. This also means that in the CTMP case the distribution in the total population (37) can only completely be estimated provided the population proportions are known and that we have samples from every subpopulation, $n_\ell > 0, \ell=1,...,L$, in order to be able to estimate all subpopulation parameters $\gamma_\ell, \ell=1,...,L$. Another difference from the CTOP situation is that in CTMP the values of Y are known before sampling, so 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 Y. So we will not consider the simultaneous probability of the response vector X_i and Y_i as in the CTOP case (38), but the conditional distribution of X_i given $Y_i=y^{(\ell)}$. The marginal likelihood to be maximized in this case can be deduced as follows. The probability of a response vector X_{i_ℓ} of a randomly sampled student from subpopulation ℓ is

$$P_{\beta,\gamma_\ell}(x_{i_\ell} \mid Y_i=y^{(\ell)}) = \int_{\theta_{i_\ell}} P_\beta(x_{i_\ell} \mid \theta_{i_\ell}).\ P_{\gamma_\ell}(\theta_{i_\ell} \mid Y_i=y^{(\ell)})\ d\theta_{i_\ell} = \int_{\theta_{i_\ell}} P_\beta(x_{i_\ell} \mid \theta_{i_\ell}).g_{\gamma_\ell}(\theta_{i_\ell})\, d\theta_{i_\ell}\ . \tag{46}$$

For all the students in stratum ℓ we have

$$\prod_{i_\ell=1}^{n_\ell} P_{\beta,\gamma_\ell}(x_{i_\ell} \mid Y_i=y^{(\ell)})\ . \tag{47}$$

So within each stratum, this (marginalizing over θ_{i_ℓ}) leads to the correct inference on β and γ_ℓ according to Rubins ignorability principle as in standard complete data MML (see (27), (28) and (29)).

If we consider the total sample, while conditioning on $\boldsymbol{Y}_i = y^{(\ell)}$, we do not apply the ignorability principle, but in fact use the design distribution in our analyses. Remember that in the presence of missing data the correct distribution of the observations, conditioned on the outcome of the design variable, is given by (see (15)):

$$\int_{u_{mis}} f_{\lambda}(u_{obs}, u_{mis} \mid m)\, du_{mis} \,. \tag{48}$$

The design distribution in this case is given by:

$$P(M_i = m_{i_\ell}) = 1 \qquad \text{if} \qquad Y_i = y^{(\ell)}\,, \tag{49}$$

with $m_{i_\ell} = (1_k, 0)$ respectively the indicator of the k items and of θ in the ℓ^{th} subpopulation. In this case, without using ignorability,

$$\int_{\theta_{i_\ell}} P_{\beta,\gamma_\ell}(x_{i_\ell}, \theta_{i_\ell} \mid M_i = m_{i_\ell})\, d\theta_{i_\ell} = \int_{\theta_{i_\ell}} P_{\beta,\gamma_\ell}(x_{i_\ell}, \theta_{i_\ell} \mid Y_i = y^{(\ell)})\, d\theta_{i_\ell}\,, \tag{50}$$

is the correct distribution to infer on. And because conditional on $\boldsymbol{Y}_i = y^{(\ell)}$ there is no dependence between the response vectors patterns over strata, the likelihood of the total sample is thus given by:

$$L(\beta, \gamma; x \mid y) = \prod_{\ell=1}^{L} \prod_{i_\ell=1}^{n_\ell} \int_{\theta_{i_\ell}} P_{\beta}(x_{i_\ell} \mid \theta_{i_\ell}) \cdot g_{\gamma_\ell}(\theta_{i_\ell})\; d\theta_{i_\ell}\,. \tag{51}$$

Again we have the product of L ordinary marginal likelihoods from which as before the parameters β and, for all strata with $n_\ell > 0$, γ_ℓ can be estimated.

In the CTMP case it is in principle also possible to make no use of the background information or subpopulation structure in calibrating the items. Defining

$$g_{\gamma^*}(\theta) = \sum_{\ell=1}^{L} \pi_\ell^*\, g_{\gamma_\ell}(\theta) = \sum_{\ell=1}^{L} (n_\ell / n) \cdot g_{\gamma_\ell}(\theta) \tag{52}$$

we find,

$$L(\beta, \gamma^*; x) = \prod_{\ell=1}^{L} \prod_{i=1}^{n_\ell} \int_{\theta_i} P_{\beta}(x_i \mid \theta_i) \cdot g_{\gamma^*}(\theta_i)\; d\theta_i\,, \tag{53}$$

which is formally identical to (45), that is the CTOP case when the background information is not used. However, this formal identity causes a lot of problems. Ignoring the background information can mean two different things. First, the subpopulation membership for each student is ignored, but the sample size n_ℓ for each subpopulation is known and used. Second, both membership and subsample sizes are ignored. In both cases $g_{\gamma^*}(\theta)$ is a finite mixture of subpopulation densities, but in the former case the weights are known ($\pi_\ell^* = n_\ell/n$) and in the latter case they have to be estimated. The estimation of $g_{\gamma^*}(\theta)$ is not easy as may be seen from a simple example. Suppose $L=2$, and $g_{\gamma_\ell}(\theta)$ is the normal probability density function. Except for the trivial case where $g_{\gamma_1} \equiv g_{\gamma_2}$, $g_{\gamma^*}(\theta)$ is not a normal probability density, and $g_{\gamma^*}(\theta)$ is to be estimated as a mixture, which may be difficult since the membership of the students is not used. Moreover, if $\pi_1^* = \pi_2^*$, it is not possible to identify the subpopulations. In case n_ℓ is not known, the estimation problems become even more severe, and identification is not possible. Replacing $g_{\gamma^*}(\theta)$ by a single normal probability density function is a specification error, yielding inconsistent estimates $g_{\gamma^*}(\theta)$ and β as well. In summary, then, it is theoretically possible not to use the background information and to obtain consistent estimates of $g_{\gamma_\ell}(\theta)$. Of course π_ℓ^* (or their estimators) only represent the sampling scheme and therefor (52) is in general an inconsistent estimator of the density of the total population.

We will extend the discussion of using or not using the background information to targeted testing. Remember that also in a targeted testing design two possible roles of the background variable Y can be distinguished. First in TTOP, Y has no role in the sampling of the students, but depending on the values of Y different subsets of the item pool are administered. Second in TTMP 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 $g_\gamma(\theta)$ (37). For students with value $y^{(\ell)}$ of Y_i denote with $\beta_{(\ell)}$ the k_ℓ -vector of the item parameters of the items administered and with r_ℓ the accompanying value of the item indicator variable (see (21)). 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 m_i effects is:

$$\left.\begin{aligned} U_{obs,i} &= (X_{obs,i},\ Y_i) \\ U_{mis,i} &= (X_{mis,i},\ \theta_i) \end{aligned}\right\},\quad (i=1,\ldots,n). \tag{54}$$

And the design distribution (compare to the CTOP case (40)) follows from (25):

$$P(M_i = (r_\ell, 1,0)) = P(R_i = r_\ell) = P(Y_i = y^{(\ell)}) \ ,\quad (i=1,\ldots,n)\ , \tag{55}$$

or

$$\left.\begin{aligned} P(M_i = (r_\ell,1,0) \mid Y_i = y^{(\ell)}) &= 1 \\ P(M_i = (r_\ell,1,0) \mid Y_i \neq y^{(\ell)}) &= 0 \end{aligned}\right\},\quad (\ell=1,\ldots,L). \tag{56}$$

From (56) 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:

$$\begin{aligned} &P_{\beta,\gamma}(x_{obs,i},\ Y_i=y^{(\ell)}) = \int_{x_{mis,i}} \int_{\theta_i} P_{\beta,\gamma}(x_{obs,i},\ x_{mis,i},\ Y_i=y^{(\ell)},\theta_i)\ d\theta_i\ dx_{mis,i} = \\ &\int_{\theta_i} P_{\beta_{(\ell)}}(x_{obs,i} \mid Y_i=y^{(\ell)},\ \theta_i).\ P_\gamma(\theta_i \mid Y_i=y^{(\ell)}).\ P(Y_i=y^{(\ell)})\ d\theta_i = \\ &\int_{\theta_i} P_{\beta_{(\ell)}}(x_{obs,i} \mid \theta_i).\ g_{\gamma_\ell}(\theta_i).\ \pi_\ell\ d\theta_i = \\ &\prod_{\ell=1}^{L} \pi_\ell^{y_{i\ell}} \cdot \prod_{\ell=1}^{L} \int_{\theta_i} \left[P_{\beta_{(\ell)}}(x_{obs,i} \mid \theta_i).\ g_{\gamma_\ell}(\theta_i)\right]^{y_{i\ell}}\ d\theta_i\ . \end{aligned} \tag{57}$$

The likelihood of the total sample is given by:

$$\begin{aligned} &L(\beta,\gamma,\pi;\ x_{obs},y) = \prod_{i=1}^{n} P_{\beta,\gamma}(x_{obs,i},\ Y_i = y^{(\ell)}) = \\ &\prod_{i=1}^{n} \prod_{\ell=1}^{L} \pi_\ell^{y_{i\ell}} \cdot \prod_{i=1}^{n} \prod_{\ell=1}^{L} \int_{\theta_i} \left[P_{\beta_{(\ell)}}\ (x_{obs,i} \mid \theta_i).\ g_{\gamma_\ell}(\theta_i) \right]^{y_{i\ell}} d\theta_i\ . \end{aligned} \tag{58}$$

The likelihood takes the same form as in the CTOP case ((38) and (39)), with the understanding that the L ordinary likelihoods do not all contain the same item parameters. The estimates of the parameters follow in the same way as described before.

If we do not use the background information in the TTOP case, the partition the observed design variable m_i establishes becomes (compare to (42) and (43)):

$$\left.\begin{aligned} U_{obs,i} &= X_{obs,i} \\ U_{mis,i} &= (X_{mis,i}, Y_i, \theta_i) \end{aligned}\right\}, \qquad (i=1,\ldots,n). \tag{59}$$

The design distribution is given by:

$$\left.\begin{aligned} &P(M_i = (r_i,0,0) \mid Y_i = y^{(\ell)}) = 1 \\ &P(M_i = (r_i,0,0) \mid Y_i \neq y^{(\ell)}) = 0 \end{aligned}\right\}, \quad (\ell=1,\ldots,L). \tag{60}$$

We see that the MAR condition in this case is not fulfilled, because the design distribution depends on values of Y_i which are considered as missing if we do not use Y in the analyses. Not using Y in the TTOP case is not justified by the ignorability principle and can lead to incorrect estimates of the parameters. For an example see Eggen (1990).

Finally, the TTMP case which has the same sampling situation as in the CTMP case. The design distribution is given by (compare to (49) and (55)):

$$P(M_i = m_{i_\ell} = (r_\ell, 0)) = 1 \quad \text{if} \quad Y_i = y^{(\ell)}. \tag{61}$$

The conditional distribution of a response vector given $Y_i = y^{(\ell)}$ is considered. So we do not use the ignorability principle but explicitly condition on the design variable in the analyses. For a randomly sampled student from subpopulation ℓ we have:

$$\begin{aligned} &P_{\beta_{(\ell)},\gamma_\ell}(x_{obs,i_\ell} \mid m_{i_\ell}) = P_{\beta_{(\ell)},\gamma_\ell}(x_{obs,i_\ell} \mid Y_i = y^{(\ell)}) = \\ &\int_{x_{mis,i_\ell}} \int_{\theta_{i_\ell}} P_{\beta_{(\ell)},\gamma_\ell}(x_{obs,i_\ell}, x_{mis,i_\ell}, \theta_{i_\ell}) \mid Y_i = y^{(\ell)}) \cdot P_{\gamma_\ell}(\theta_{i_\ell} \mid Y_i = y^{(\ell)}) \, d\theta_{i_\ell} \, dx_{mis,i} = \\ &\int_{\theta_{i_\ell}} P_{\beta_{(\ell)}}(x_{obs,i_\ell} \mid \theta_{i_\ell}) \cdot g_{\gamma_\ell}(\theta_{i_\ell}) \, d\theta_{i_\ell} \, . \end{aligned} \tag{62}$$

And for the total sample we have the likelihood

$$\prod_{\ell=1}^{L} \prod_{i_\ell=1}^{n_\ell} \int_{\theta_{i_\ell}} P_{\beta(\ell)}(x_{obs,i_\ell} \mid \theta_{i_\ell}) . g_{\gamma_\ell}(\theta_{i_\ell}) \; d\theta_{i_\ell} \; . \tag{63}$$

As before the parameters β and γ_ℓ, $\ell=1,...,L$ (provided $n_\ell>0$) can be estimated from (63).

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 $g_{\gamma^*}(\theta)$ (see (52) in the CTMP case) then we are in fact in the TTOP situation for which it was shown ((59) and (60)) that by ignoring Y the MAR condition for ignorability is not fulfilled.

In Table 1 an overview of the preceding results is presented.

Table 1. Overview of MML estimation in the presence of background information.

Design	β	γ_ℓ	π_ℓ	γ
CTOP				
Using Y	C	C	C	
Ignoring Y	C	NA	NA	C
CTMP				
Using Y	C	C	C	
Ignoring Y (one pop)	-	NA	NA	-
Ignoring Y (mix pop)	C^-	C^-	C^-	
TTOP				
Using Y	C	C	C	
Ignoring Y	-	-	-	
TTMP				
Using Y	C	C	C	
Ignoring Y	-	-	-	

In the second column of Table 1 it is indicated whether using a MML procedure for item calibration leads to correct estimates of the parameters according to the ignorability principle. The parameters are defined as before, with the understanding that γ stands for the

parameters of one population distribution without regarding the subpopulation structure as in (37). C stands for correct estimates, C^- for in principle correct, but practically almost infeasible to get the estimates, NA is for not available and finally - is for not correct.

In CTOP complete testing we see that 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, there are differences in the available estimates of the ability distributions. In CTMP only using $\boldsymbol{Y}$ guarantees correct estimates. Mislevy and Sheenan (1989) showed that using $\boldsymbol{Y}$ is more useful if the positive relation between $\boldsymbol{Y}$ and θ is stronger. But then the assumption of one population distribution is also less realistic. In practice, however, assuming the finite mixture ability distribution (37) or (52) asks for facilities in computer programs which are not always available. For example, BILOG has no facilities to estimate more than one ability distribution and to estimate the proportions. Using $\boldsymbol{Y}$ is not possible in this program and if we do not use $\boldsymbol{Y}$, estimating γ necessarily as the parameters of one (normal) distribution can only be justified when all subpopulations have the same distribution. The computer program OPLM does has facilities to estimate more than one normal ability distribution.

However, in incomplete targeted testing designs, TTOP as well as TTMP, there is no choice whether the background information $\boldsymbol{Y}$ must be used. Ignoring $\boldsymbol{Y}$ never leads to correct inferences on the item parameters or the population parameters. This is easily understood since by ignoring $\boldsymbol{Y}$, $\boldsymbol{Y}$ is part of the missing data and the condition of MAR will never be fulfilled. 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. Because standard computer implementation of MML procedures (e.g., in BILOG) have no facilities to use $\boldsymbol{Y}$, and always assume one random sample of students, in practice, many failures are made.

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, except in one situation, useful. This situation is, when we are sampling from multiple populations where 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 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 this ignorability principle is also useful in CML estimation. In using the ignorability principle we fix the random design variable $\boldsymbol{M}$ at the observed pattern of missing data $\boldsymbol{m}$ and assume that the values $\boldsymbol{u}_{obs}$ are realizations of the marginal distribution of $\boldsymbol{U}_{obs}$ (14):

$$\int_{u_{mis}} f(u_{obs}, u_{mis})\, du_{mis} \quad . \tag{64}$$

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

$$\int_{u_{mis}} f(u_{obs}, u_{mis} \mid m) du_{mis} \quad , \tag{65}$$

the conditional distribution of $\boldsymbol{U}_{obs}$ given $\boldsymbol{M}=\boldsymbol{m}$, is not used in the analysis, but only the marginal distribution of the observed responses. Note that in the CML case, the design variable $\boldsymbol{M}_i$ and the item indicator variable $\boldsymbol{R}_i$ are the same because the only variables inferred on are the item responses $\boldsymbol{X}$, and θ is not treated as a random variable as in MML.

It will be clear that ignoring the design in CML estimation is only justified (consistency and asymptotic efficiency of the estimators) if for an individual observed response vector $\boldsymbol{X}_{obs,i}$ there exists a sufficient statistic $\boldsymbol{S}_{obs,i} = \boldsymbol{S}(\boldsymbol{X}_{obs,i})$ for θ_i in the marginal distribution (64) and the distribution of $\boldsymbol{S}_{obs,i}$ is at least S-ancillary with respect to $\boldsymbol{\beta}$. It can easily be shown that in the IRT models we consider, for example in the Rasch model the sumscore

$$\boldsymbol{S}_{obs,i} = \sum_{j \in obs} \boldsymbol{X}_{ij} \quad , \tag{66}$$

is not only <u>not</u> sufficient for θ_i in the marginal distribution of the observations $\boldsymbol{X}_{obs,i}$, but also not sufficient in the distribution of all observed data $(\boldsymbol{X}_{obs,i}, \boldsymbol{R}_i)$. $\boldsymbol{S}_{obs,i}$ is only sufficient in the conditional distribution of the responses given the item indicator variable $\boldsymbol{r}_i$. 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 $(0<\phi<1)$:

$$P(\boldsymbol{R}_i=\boldsymbol{r}_1=(1,1,0)) = \phi, \quad \text{and} \quad P(\boldsymbol{R}_i=\boldsymbol{r}_2=(1,0,1)) = 1-\phi\,. \tag{67}$$

In Table 2 the relevant probabilities for all outcomes with $S_{obs}=1$, with $\exp(\theta)=\xi$, are given.

Table 2. Probabilities for all outcomes with $S_{obs}=1$.

x_{obs},r	$P(x_{obs},r)$ (1)	$P(x_{obs} \mid r_1)$ (2)	$P(x_{obs} \mid r_2)$ (3)
10,110	$\frac{\phi.\xi\epsilon_1}{(1+\xi\epsilon_1)(1+\xi\epsilon_2)}$	$\frac{\xi\epsilon_1}{(1+\xi\epsilon_1)(1+\xi\epsilon_2)}$	0
01,110	$\frac{\xi\epsilon_2}{(1+\xi\epsilon_1)(1+\xi\epsilon_2)}$	$\frac{\xi\epsilon_2}{(1+\xi\epsilon_1)(1+\xi\epsilon_2)}$	0
10,101	$\frac{(1-\phi).\xi\epsilon_1}{(1+\xi\epsilon_1)(1+\xi\epsilon_3)}$	0	$\frac{\xi\epsilon_1}{(1+\xi\epsilon_1)(1+\xi\epsilon_3)}$
01,101	$\frac{(1-\phi).\xi\epsilon_3}{(1+\xi\epsilon_1)(1+\xi\epsilon_3)}$	0	$\frac{\xi\epsilon_3}{(1+\xi\epsilon_1)(1+\xi\epsilon_3)}$
1	$\frac{\phi.\xi(\epsilon_1+\epsilon_2)}{(1+\xi\epsilon_1)(1+\xi\epsilon_2)}+\frac{(1-\phi).\xi(\epsilon_1+\epsilon_3)}{(1+\xi\epsilon_1)(1+\xi\epsilon_3)}$	$\frac{\xi(\epsilon_1+\epsilon_2)}{(1+\xi\epsilon_1)(1+\xi\epsilon_2)}$	$\frac{\xi(\epsilon_1+\epsilon_3)}{(1+\xi\epsilon_1)(1+\xi\epsilon_3)}$
s_{obs}	$P(s_{obs})$	$P(s_{obs} \mid r_1)$	$P(s_{obs} \mid r_2)$

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

$$P_{\theta,\beta,\phi}(\boldsymbol{x}_{obs},\boldsymbol{r}) = \prod_i P_{\theta_i,\beta,\phi}(\boldsymbol{x}_{obs,i} \mid \boldsymbol{r}_i)\ .\ P_{\phi}(\boldsymbol{r}_i)\ . \tag{68}$$

We use the same notation as before. We distinguish T values of the design variable $r_t, t=1,...,T$; 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 (68) as:

$$P_{\theta,\beta,\phi}(x_{obs},r) = \prod_{t=1}^{T} \prod_{i=1}^{n_t} P_{\theta_i,\beta_{(t)},\phi}(x_{obs,i} \mid r_t) \cdot P_{\phi}(r_t) \quad . \tag{69}$$

We see in (69) that we have in fact the product of T complete data likelihoods. For every t the first factor in the right-hand side of (69) can, as in complete data CML (see (5)), be written as

$$\prod_{i=1}^{n_t} P_{\theta_i,\beta_{(t)},\phi}(x_{obs,i} \mid r_t) = \prod_{i=1}^{n_t} P_{\beta_{(t)}}(x_{obs,i} \mid s_{obs,i}, r_t) \cdot P_{\theta_i,\beta_{(t)},\phi}(s_{obs,i} \mid r_t) \, . \tag{70}$$

And the first factor in the right-hand side of (70) is again free of any incidental parameters, and

$$L_c = \prod_{t=1}^{T} \prod_{i=1}^{n_t} P_{\beta_{(t)}}(x_{obs,i} \mid s_{obs,i}, r_t) \tag{71}$$

can be used for CML estimation of β. 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 ignorability principle 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 (71) depends of course again on the properties of the part of the total likelihood (69) we neglect in that case. The neglected part in CML is (combining (69), (70) and (71))

$$\prod_{t=1}^{T} \prod_{i=1}^{n_t} P_{\theta_i,\beta_{(t)},\phi}(s_{obs,i}, r_t) = \prod_{t=1}^{T} \prod_{i=1}^{n_t} P_{\theta_i,\beta_{(t)},\phi}(s_{obs,i} \mid r_t) \cdot \prod_{t=1}^{T} \prod_{i=1}^{n_t} P_{\phi}(r_t) \, . \tag{72}$$

We will discuss the properties of (72) for the three considered design types next.

CML in Random Incomplete Designs

In random incomplete designs the design distribution is given by (22). Considering the first factor of the part of the likelihood we neglect in CML (72), we see this factor consists of the product of T complete data distributions of the sufficient statistics S_{obs}. And because every term in this product is S-ancillary with respect to the item parameters $\boldsymbol{\beta}$, the product is also S-ancillary. From the design distribution (22) it is easily seen that the second part of (72), $P_{\phi}(\boldsymbol{r}_t)$, does not depend on the item parameters at all. Which means that this distribution is ancillary with respect to $\boldsymbol{\beta}$. As a consequence, (72) is S-ancillary with respect to $\boldsymbol{\beta}$, which 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 (72) is S-ancillary with respect to the item parameters $\boldsymbol{\beta}$ for the same reasons 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 (23) we can write the second part as:

$$\prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\phi}(\boldsymbol{R}_i=\boldsymbol{r}_t) = \prod_{t=1}^{T}\prod_{i=1}^{n_t} P(\boldsymbol{R}_i=\boldsymbol{r}_t \mid \boldsymbol{x}_{obs,i}\in C_t)\,.\, P_{\boldsymbol{\beta}_{(obs)},\theta_i}(\boldsymbol{x}_{obs,i}\in C_t) = \prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\boldsymbol{\beta}_{(obs)},\theta_i}(\boldsymbol{x}_{obs,i}\in C_t)\,. \tag{73}$$

We see that (73) is for every t directly dependent of the item parameters of the items used for establishing the design. This means that (72), which would be neglected in CML estimation, cannot be S-ancillary with respect to $\boldsymbol{\beta}$. 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 (see also Glas (1988)).

CML in Targeted Testing Designs

In targeted testing designs the value of a background variable $\boldsymbol{Y}$ determines the design. The design distribution is given by (25) and (26). Before we made the distinction between the two sampling roles $\boldsymbol{Y}$ can play in the design and using or not using $\boldsymbol{Y}$ 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 $\boldsymbol{X}_i$ and of $\boldsymbol{Y}_i$ of student i is given by

$$P_{\theta_i,\beta,\pi_\ell}(\boldsymbol{x}_i, \boldsymbol{Y}_i=y^{(\ell)}) = P_{\theta_i,\beta}(\boldsymbol{x}_i \mid \boldsymbol{Y}_i=y^{(\ell)}).P_{\pi_\ell}(\boldsymbol{Y}_i=y^{(\ell)}) . \tag{74}$$

Conditioning on the sufficient statistic $\boldsymbol{S}_i$ gives:

$$\begin{aligned} P_{\theta_i,\beta,\pi_\ell}(\boldsymbol{x}_i,\boldsymbol{Y}_i=y^{(\ell)}) &= P_{\theta_i,\beta}(\boldsymbol{x}_i \mid s_i,\boldsymbol{Y}_i=y^{(\ell)}).P_{\theta_i,\beta}(s_i \mid \boldsymbol{Y}_i=y^{(\ell)}).P_{\pi_\ell}(\boldsymbol{Y}_i=y^{(\ell)}) \\ &= P_{\beta}(\boldsymbol{x}_i \mid s_i).P_{\theta_i,\beta}(s_i \mid \boldsymbol{Y}_i=y^{(\ell)}).P_{\pi_\ell}(\boldsymbol{Y}_i=y^{(\ell)}) . \end{aligned} \tag{75}$$

In (75) $\boldsymbol{Y}_i=y^{(\ell)}$ cancels in $P_{\beta}(\boldsymbol{x}_i \mid s_i)$ because of the local independence of the item responses. The complete likelihood of the sample is given by:

$$\prod_i P_{\beta}(\boldsymbol{x}_i \mid s_i) \cdot \prod_i P_{\theta_i,\beta}(s_i \mid \boldsymbol{Y}_i=y^{(\ell)}) \cdot P_{\pi_\ell}(\boldsymbol{Y}_i=y^{(\ell)}) . \tag{76}$$

From (76), the first factor is used in CML estimation. As before, the first part of the second factor is S-ancillary with respect to $\boldsymbol{\beta}$ and the second part is independent of it. So CML is a justified procedure to estimate $\boldsymbol{\beta}$. 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 $P_{\theta_i,\beta}(\boldsymbol{x}_i \mid \boldsymbol{Y}_i=y^{(\ell)})$ with as a consequence that $P_{\pi_\ell}(\boldsymbol{Y}_i=y^{(\ell)})$ cancels in (75) and (76). 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.

Secondly we consider incomplete targeted testing. Here we distinguish as many values (L) of the design variable r_i as we distinguish values of the background variable Y_i. If we rewrite the total likelihood as before ((69), (71) and (72)) we see that the conditional likelihood to be maximized is:

$$\prod_{\ell=1}^{L}\prod_{i=1}^{n_\ell} P_{\beta_{(\ell)}}(x_{obs,i} \mid s_{obs,i}, r_\ell, Y_i=y_{(\ell)}), \tag{77}$$

and the neglected part becomes

$$\begin{aligned}&\prod_{\ell=1}^{L}\prod_{i=1}^{n_\ell} P_{\theta_i,\beta_{(\ell)},\phi,\pi}(s_{obs,i}, r_\ell, Y_i=y^{(\ell)}) = \\ &\prod_{\ell=1}^{L}\prod_{i=1}^{n_\ell} P_{\theta_i,\beta_{(\ell)},\phi,\pi}(s_{obs,i} \mid r_\ell, Y_i=y^{(\ell)}) \cdot \prod_{\ell=1}^{L}\prod_{i=1}^{n_\ell} P_{\phi,\pi}(r_\ell, Y_i=y^{(\ell)}).\end{aligned} \tag{78}$$

From the design distribution (25) it is seen that in targeted testing the events $\{R_i=r_\ell\}$ and $\{Y_i=y^{(\ell)}\}$ coincide. And the pair $\{R_i=r_\ell, Y_i=y_{(\ell)}\}$ in (77) and (78) can be replaced by either one of the two. And it can then easily be checked as before that the first part of (78) is S-ancillary with respect to β and the second part independent of it. So CML estimation, on the basis of the conditional likelihood (77), is justified in targeted testing.

Conclusion

In the preceding it was shown for the three most common random 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, resulting in consistent and asymptotic efficient estimators, was shown to be the S-ancillarity of the neglected part of the total likelihood of the data. Summarized it can be said that in random incomplete designs but MML and CML are possible. In multistage testing designs only MML is eligible for item calibration, CML is not justified. In targeted

testing CML is always possible, while MML is only justified when there are as many marginal ability distributions specified as designs (or strata in complete testing).

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 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 the paper 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 polytimously scored items and for models with more than one item parameter.

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.

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. (1980). Applications of item response theory to practical testing problems. Hilsdale: Lawrence Erlbaum Associates.

Mislevy, R.J. & Bock, R.D. (1982). *BILOG: Maximum likelihood item analysis and test scoring with logistic models for binary items.* Chicago: International Educational Services.

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. (1992). Het eenparameter logistisch model (OPLM). Een theoretische inleiding en een hanleiding bij het computerprogramma. [The generalized one-parameter logististic model (OPLM). A theoretical introduction and a manual of the computer program.] *OPD memorandum 92-3*. 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.

Recent Measurement and Research Department Reports:

91-1 N.D. Verhelst & N.H. Veldhuijzen. A New Algorithm for Computing Elementary Symmetric Functions and Their First and Second Derivatives.

91-2 C.A.W. Glas. Testing Rasch Models for Polytomous Items: With an Example Concerning Detection of Item Bias.

91-3 C.A.W. Glas & N.D. Verhelst. Using the Rasch Model for Dichotomous Data for Analyzing Polytomous Responses.

91-4 N.D. Verhelst & C.A.W. Glas. A Dynamic Generalization of the Rasch Model.

91-5 N.D. Verhelst & H.H.F.M. Verstralen. The Partial Credit Model with Non-Sequential Solution Strategies.

91-6 H.H.F.M. Verstralen & N.D. Verhelst. The Sample Strategy of a Test Information Function in Computerized Test Design.

91-7 H.H.F.M. Verstralen & N.D. Verhelst. Decision Accuracy in IRT Models.

91-8 P.F. Sanders & T.J.H.M. Eggen. The Optimum Allocation of Measurements in Balanced Random Effects Models.

91-9 P.F. Sanders. Alternative Solutions for Optimization Problems in Generalizability Theory.

91-10 N.D. Verhelst, H.H.F.M. Verstralen & T.J.H.M. Eggen. Finding Starting Values for the Item Parameters and Suitable Discrimination Indices in the One-Parameter Logistic Model.

92-1 N.D. Verhelst, H.H.F.M. Verstralen & M.G.H. Jansen. A Logistic Model for Time Limit Tests.

92-2 F.H. Kamphuis. Estimation and Prediction of Individual Ability in Longitudinal Studies.