Psicológica (2011), 32, 107-132.

Item calibration in incomplete testing designs

Theo J.H.M. Eggen* & Norman D. Verhelst**

**Cito/University of Twente, The Netherlands*

***Cito, The Netherlands*

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 Sheenan (1989) have shown that in incomplete 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 and extended for more situations. 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. The problems with incomplete designs are not generally recognized in practical situations. This is due to the stochastic nature of the incomplete designs which is not taken into account in standard computer algorithms. For that reason, incorrect uses of standard MML- and CML-algorithms are discussed.

Introduction

Within the framework of item response theory (IRT) item calibration involves the estimation of the item parameters in the chosen IRT model. For these so-called scaling procedures often data gathered in incomplete designs are used. In item banking 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. In equating studies, the incomplete designs is mostly a starting point, because only partly overlapping tests are administered to different groups of students.

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

* E-mail: theo.eggen@cito.nl

problems with incomplete designs are not generally recognized. This is due to the ignorance of the consequences of the stochastic nature of the incomplete designs which is not taken into account in these computer algorithms. In particular this is the case in equating studies where item calibration in incomplete designs as studied here is often called concurrent calibration and is then compared with linking on the same scale separately calibrated tests with data from complete designs. (see e.g., Hanson & Béguin, 2002).

In this study (concurrent) calibration procedures in incomplete testing designs are reviewed. The statistical approach of use imputation techniques (Little & Rubin, 1987) in the handling of missing data and subsequently analysing complete data will not be considered in this study. Here, the likelihood approach, in which as well observed as missing data are modelled, will be studied. The justification of applying MML and CML procedures in the incomplete designs will be studied.

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 (1996), 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 in this study are partly recapitulations of their work and are extended to other situations. The results for the justification of CML estimation of the item parameters in incomplete designs are necessarily deduced via a different approach.

Item Response Theory

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)

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

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.

By the usual assumptions of (local) independence the probability of the response pattern is given by (with $\boldsymbol{\theta} = (\theta_i), i = 1,...,n$ and $\boldsymbol{\beta} = (\beta_j), j = 1,...,n$)

$$P_{\boldsymbol{\theta},\boldsymbol{\beta}}(x) = \prod_i P_{\theta_i,\boldsymbol{\beta}} = \prod_i \prod_j P_{\theta_i,\beta_j}(x_{ij}) . \qquad (2)$$

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 observed response pattern $\boldsymbol{X} = \boldsymbol{x}$, or the likelihood function

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

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 Estimation

If it is possible to construct a sufficient statistic $S(X_i)$ 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

$$P_{\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{x}) = \prod_i P_{\boldsymbol{\beta}}(x_i \mid s(x_i)).P_{\theta_i,\boldsymbol{\beta}}(s(x_i)) , \qquad (3)$$

In (3) $P_{\theta_i,\boldsymbol{\beta}}(s(x_i))$ is the distribution of the sufficient statistic $S(X_i), i = 1,...,n$ And the first factor $\prod_i P_{\boldsymbol{\beta}}(x_i \mid s(x_i))$, is the simultaneous conditional probability of the observed responses $\boldsymbol{x}$, which does not depend on the ability parameters because of the sufficiency of $S(X_i)$ for θ_i. In CML estimation we then proceed estimating the item parameters by just maximizing this conditional likelihood function with respect to $\boldsymbol{\beta}$:

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

In CML estimation of the item parameters only random variations of the observations, fixing (given) the values of the conditioning statistics $s(x_i)$ 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 (3). If the distribution of the sufficient statistic $s(x_i)$ 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 β 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 $s(x)$, is very small already at short test lengths. 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 neglected.

Marginal Maximum Likelihood Estimation

In MML estimation, model (2) is extended by assuming that the ability parameters θ_i are a random sample from a population with probability density function given by $g_\gamma(\theta)$, with $\boldsymbol{\gamma}$ the (possibly vector valued) parameter of the ability distribution. Thus the response pattern X 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 $\boldsymbol{X}$,

$$P_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{x}) = \int P_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{x},\theta)d\theta = \prod_i \int P_{\boldsymbol{\beta}}(x_i|\theta_i)g_{\boldsymbol{\gamma}}(\theta_i)d\theta_i, \tag{4}$$

where $P_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\boldsymbol{x},\theta)$ is the simultaneous distribution of the response pattern $\boldsymbol{X}$ and the ability $\boldsymbol{\theta}$. $P_{\boldsymbol{\beta}}(\boldsymbol{x}_i|\theta_i) = \prod_i \int P_{\beta_j}(x_{ij}|\theta_i)$ is the IRT model as in (2), giving the probability of a response vector i of person , with ability θ_i.

In MML estimation the item parameters β 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,

$$L_m(\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{x}) = \prod_i \int P_{\boldsymbol{\beta}}(x_i|\theta_i)g_{\boldsymbol{\gamma}}(\theta_i)d\theta_i. \tag{5}$$

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 $\boldsymbol{\gamma} = (\mu,\sigma^2)$. Bock and Aitkin (1981) were the first to give computational procedures for maximizing (5) 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 $\boldsymbol{U} = (U_1,...,U_m)$ be a vector random variable with probability density function $f_\tau(\boldsymbol{u})$, with a vector parameter τ, on which we want to draw inferences on the basis of the data, a sample realization $\boldsymbol{u}$. Assume for convenience that $m = n \cdot k$, 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 of its observed value: $\boldsymbol{U} = (\boldsymbol{U}_{obs}, \boldsymbol{U}_{mis})$ and $\boldsymbol{u} = (\boldsymbol{u}_{obs}, \boldsymbol{u}_{mis})$. 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:

$$\mathrm{P}_\varphi(\boldsymbol{M} = \boldsymbol{m} \,|\, \boldsymbol{U} = \boldsymbol{u}) = h_\varphi(\boldsymbol{m} \,|\, \boldsymbol{u}),$$

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 can be dependent on the parameter of interest τ: 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 $\boldsymbol{M}$ and $\boldsymbol{U}_{obs}$ and we want to infer on the parameter τ of the distribution of the only partially observed $\boldsymbol{U}$. In the presence of missing data, the basis for inference on τ should be the joint distribution of $\boldsymbol{M}$ and $\boldsymbol{U}_{obs}$:

$$\int_{u_{mis}} f_{\tau,\varphi}(\boldsymbol{u}, \boldsymbol{m}) d\boldsymbol{u}_{mis} = \int_{\boldsymbol{u}_{mis}} f_\tau(\boldsymbol{u}) \cdot h_\varphi(\boldsymbol{m} \,|\, \boldsymbol{u}) d\boldsymbol{u}_{mis}. \tag{6}$$

Because we are only interested to infer on the parameter τ of the distribution of the partially observed $\boldsymbol{U}$, 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 $\boldsymbol{M}$ at the observed pattern of missing data m and (b) assuming that the values of the observed $\boldsymbol{U}_{obs}$ data are realizations of the marginal density of $\boldsymbol{U}_{obs}$:

$$f_\tau(\boldsymbol{u}_{obs}) = \int_{u_{mis}} f_{\tau,\varphi}(\boldsymbol{u}, \boldsymbol{m}) du_{mis}. \tag{7}$$

When we ignore the process that causes the missing data, not all possible random variation in the data due to sampling of $\boldsymbol{M}$ and $\boldsymbol{U}_{obs}$ is considered, but only random variation due to $\boldsymbol{U}_{obs}$ fixing the random variable $\boldsymbol{M}$ at the particularly observed pattern m. The generally more convenient form (7) is used instead of (6) 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 influence of φ on τ: possible restrictions, due to $\varphi\ \theta$, are not taken in account in the inference on τ. Secondly, it is understood that the data u_{obs} are in fact no realizations of (7) but of the conditional density of U_{obs} given the random variable M took the fixed value m:

$$\int f_{\tau,\varphi}(u \mid m)du_{mis} = \int \frac{f_{\tau,\varphi}(u,m)}{f_{\varphi}(m)} du_{mis} = \int \frac{f_{\tau}(u).h_{\varphi}(m \mid u)}{\int f_{\tau}(u).h_{\varphi}(u \mid m)d\,u} du_{mis}\,, \qquad (8)$$

which is in general not equal to (7).

We now specify Rubin's (1976) sufficient conditions under which ignoring the process that causes the missing data yields the correct direct likelihood inference about τ. 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 $h_{\varphi}(m \mid u)$. Define:

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

$h_{\varphi}(m \mid u_{obs}, u_{mis}) = h_{\varphi}(m \mid u_{obs})$ for all u_{mis},

that is, the missingness of the data does not depend on the not observed 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

$h_{\varphi}(m \mid u_{obs}, u_{mis}) = h_{\varphi}(m)$ for all u_{mis} and u_{obs}.

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 instead of using the full-likelihood

$$L(\tau,\varphi;u_{obs},m)=f_{\tau,\varphi}(u_{obs},m)=\int_{u_{mis}} f_{\tau,\varphi}(u,m)du_{mis}, \tag{9}$$

the simple likelihood function

$$L(\tau;u_{obs})=f_{\tau}(u_{obs})=\int_{u_{mis}} f_{\tau}(u)du_{mis} \tag{10}$$

can be used for inferring on τ. 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 (9) or (10) will have the same statistical properties, such as consistency or asymptotic normality. In general, then stronger conditions have to be fulfilled (Rubin, 1976).

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 to describe 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, 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 R_i can take T values:

$$\mathbf{r}_t=perm_t(1_{k_t},0_{k-k_t}),\ (t=1,...,T). \tag{11}$$

Each value of the design vector $\mathbf{r}_t$ is a permutation of the vector $(1_{k_t},0_{k-k_t})$, 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 M is strongly related to the item indicator R. In our applications it is always true that $R \subset M$. But R concerns only the indication whether items are observed, while M also concerns the observation or missingness of other variables considered in a problem. More specifically, when the ability θ is considered as a random variable as in MML estimation (5), we will use the indicator variable M, having a value zero for all realizations of θ.

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 φ_t such that $\sum_{t=1}^{T} \varphi_t = 1$. The distribution of the item indicator variable R_i is given by:

$$\mathrm{P}(R_i = r_t) = \varphi_t \text{ with } (t = 1,...,T),(i = 1,...,n). \tag{12}$$

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 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.

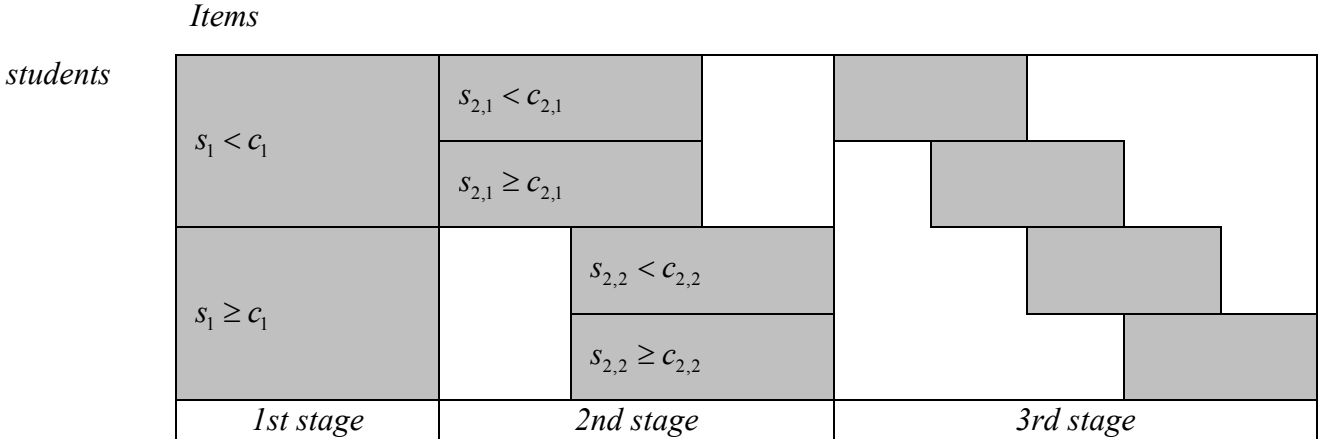

Figure 1. Example of a multistage testing design.

In Figure 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, 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 R_1 can take as many values as there are tests T (see (11)). The distribution of R_1 has always the following form:

$$\mathrm{P}(R_i = r_t \mid x_{obs,i}) = \varphi_t(x_{obs,i}), \text{ with } (t = 1,...,T), (i = 1,...,n). \tag{13}$$

If a function of observed item scores $\boldsymbol{x}_{obs,i}$ meets a criterion for getting test t, the item indicator variable R_1 takes the value r_t with probability φ_t. If the criterion is not met the probability is 1-φ_t. It should be understood that in a multistage testing design the probability of a certain design is not constant for all values of $x_{obs,i}$, because in that case the design is random incomplete.

Example 1.

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 administered. 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 $\varphi_t(x_{obs,i})$ (13) can then be defined as: $\varphi_1(x_{obs,i}) = 1$ if $\sum_{j=1}^{3} x_{ij} \leq 1$, and $\varphi_2(x_{obs,i}) = 1$ if $\sum_{j=1}^{3} x_{ij} > 1$, where test 1 is the easier test and test 2 the harder test.

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 Y of the students. This background variable is usually positively related to the ability. Students with values of Y which are expected to have lower abilities are administered easier test forms, and students with values of 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 Y of the students is categorical (or categorized), taking (or distinguishing) T values: $y_{1,},...,y_T$. In targeted testing, for each value of Y a different subset from the total item pool is administered to the students. The variable 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 Y is only used in the assignment of items or tests to students and not in the sampling of students. Second, the 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 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 Y itself. There is, for instance, no interest to have estimates of the parameters of the ability distribution for each distinguished level of Here the students are sampled from one population with no regard to the values of Y.

In the second situation, the background variable also plays a role in sampling the students. In this case there is an explicit interest in the variable itself. A situation often occurring is that 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 the first situation, the sampled students can in general not be considered to be a sample from one population but are samples from a total population divided in subpopulations of interest.

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

In targeted designs the item indicator variable R_i for every student can again take as many values as there are tests. The distribution of R_i, is given by

$$\mathrm{P}(R_i = r_t \mid y_i = y_t) = \varphi_t(y_i),\ (t = 1,...,T). \tag{14}$$

For any (distinguished) value of the background variable Y here is a fixed probability that a certain test is administered. An example is the gender of the student. A boy $y_i = 1$ then gets with a probability $\varphi_1(y_i = 1)$ test 1 and a girl with probability $\varphi_1(y_i = 2)$. Similar probabilities can be specified for a second test. In practice, often $\varphi_t = 1,$ which means that given the value y_i, a specific test is administered. The formal resemblance between a targeted testing (14) and a multistage testing design (14) is noted. But the difference is also clear: In a targeted testing design 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 (14).

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 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 former section. 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 $\boldsymbol{U}_{obs,i}$ and $\boldsymbol{U}_{mis,i}$, $\boldsymbol{U} = (\boldsymbol{U}_{obs}, \boldsymbol{U}_{mis})$ 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 $\boldsymbol{M} = (M_1,...,M_n)$, in which every element M_i is a vector of the same length as there are variables (observed and unobserved).

The Marginal Model and Missing Data

Using the same approach as Mislevy and Sheenan (1989), the ignorability conditions for the design variable in incomplete designs for MML item parameter estimation can be checked. 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 (4) and (5), 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 $U = (X, \theta) = (X_1, \theta_1, ..., X_n, \theta_n)$ in which X_i is as before the random answer vector of student i on the k items administered. The parameter to be estimated is $\tau = (\beta, \gamma)$. In the complete data situation the 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

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

The partition which the observed design variable m_i effects is

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

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

$$\int_{u_{mis}} f_\tau(u)du_{mis} = \int P_{\beta,\gamma}(x,\theta)d\theta = \prod_i \int P_\beta(x_i|\theta_i)g_\gamma(\theta_i)d\theta_i .$$

Which is identical to (5).

In random incomplete designs and multistage testing designs the ignorability conditions are also fulfilled. In Table 1 we give for these designs and for the complete testing design respectively the observed and unobserved variables and the design distribution.

The design distribution in random incomplete and in multistage testing design follow respectively from (12) and (14). 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 $(U_{obs,i}, M)$ needed for the full likelihood:

$$\int_{u_{mis}} P_{\tau,\varphi}(u,m)du_{mis} = \int_\theta \int_{x_{mis}} P_{\beta,\gamma,\varphi}(x_{obs},x_{mis},\theta,m)dx_{mis}d\theta =$$

$$\int_\theta \int_{x_{mis}} P_{\beta,\gamma}(x_{obs},x_{mis},\theta).h_\varphi(m \mid x_{obs},x_{mis},\theta)dx_{mis}d\theta =$$

$$h_\varphi(m \mid x_{obs}) \int_\theta P_{\beta,\gamma}(x_{obs},\theta)d\theta = \qquad (15)$$

$$\prod_i h_\varphi(m_i \mid x_{obs,i}) \prod_i \int_{\theta_i} P_\beta(X_{obs,i},\theta_i).g_\gamma(\theta_i)d\theta_i .$$

Table 1. Variables in incomplete testing designs

Design	$U_{obs,i}$	$U_{mis,i}$	$h_\varphi(m_i \mid U_{obs,i}, U_{mis,i})$
complete	X_i	θ_i	$\mathrm{P}(M_i = (1_k,0)) = \mathrm{P}(R_i = (1_k)) = 1$
random incomplete	$X_{obs,i}$	$X_{mis,i}, \theta_i$	$\mathrm{P}(M_i = (r_t,0)) = \mathrm{P}(R_i = (r_t)) = \varphi_t$
multistage	$X_{obs,i}$	$X_{mis,i}, \theta_i$	$\mathrm{P}(M_i = (r_t,0) \mid X_{obs,i}) = \mathrm{P}(R_i = r_t \mid X_{obs,i}) = \varphi_t(X_{obs,i})$

In (15) 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.

Note that if we indicate by n_t the number of students taking test t with $\Sigma_{i=1}^{T} n_t = n$ and define $\beta_{(t)}$ as the k_t-vector of the item parameters of the items in test we can rewrite the second factor of (15) as

$$\prod_{i=1}^{n} \int_{\theta_i} P_{\beta}(x_{obs,i} \mid \theta_i).g_{\gamma}(\theta_i)d\theta_i = \prod_{t=1}^{T}\prod_{i=1}^{n_t} \int_{\theta_i} P_{\beta_{(t)}}(x_{obs,i} \mid \theta_i).g_{\gamma}(\theta_i)d\theta_i .$$

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) have presented a general discussion on the effect of using or not using (ignoring) the background information of the students in MML item calibration. In complete testing designs they consider the two different roles of the background variable Y can play in the sampling: students can be sampled from one population, or (stratified) from multiple subpopulations. In targeted testing designs also the same two roles of Y can be distinguished. As mentioned before 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 and in TTMP Y has both a role in the sampling of the students and in the assignment of items to students. Mislevy and Sheenan (1989) have only considered the latter situation. Their results will be summarized and will be compared and completed with the results in the TTOP situation.

Assume 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 Y for student i is defined as $y_i = (y_{il},...,y_{iL})$ 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 $y_i = y^{(\ell)}$. The ability distribution $g_{\gamma}(\theta)$ of the total population in this case can be expressed as a finite mixture of L subpopulation ability distributions:

$$g_{\gamma}(\theta) = \sum_{\ell=1}^{L} P(\theta, Y = y^{(\ell)}) = \sum_{\ell=1}^{L} P(\theta | Y = y^{(\ell)}).\mathrm{P}(Y = y^{(\ell)}) = \sum_{\ell=1}^{L} g_{\gamma_\ell}(\theta).\pi_\ell , \qquad (16)$$

in which y_ℓ 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.

In complete testing designs using or not using Y in MML item calibration is equivalent with considering Y as observed or missing data. In Mislevy and Sheenan (1989) checks of Rubins ignorability conditions in this situation are given. Summarized the results are: Y using in MML item calibration makes it possible, independent of the sampling role, to estimate the item parameters $\boldsymbol{\beta} = (\beta_1,...,\beta_k)$ and the population parameters $\gamma = (\gamma_1,..,\gamma_\ell,..,\gamma_L,\pi_1,..,\pi_\ell,..,\pi_L)$ simultaneously. The

justifiability of ignoring Y depends on the sampling role of 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 Y may lead to wrong estimates.

In targeted testing designs we first consider the TTOP situation. In TTOP we have a random sample from the total population with ability distribution $g_\gamma(\theta)$ (16). For students with value $y^{(\ell)}$ of Y_i denote with $\boldsymbol{\beta}_{(\ell)}$ 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 (11)). 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:

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

and the distribution of the missing data indicator follows from (14):

$$\mathrm{P}(M_i = (r_\ell,1,0) \mid Y_i = y^{(\ell)}) = \varphi_\ell, (\ell = 1,..., \mathrm{L}) . \tag{17}$$

Note that the design vector M_i variable has one element more compared to the situations in complete, in random incomplete and in multistage testing indicating the observation of Y_i. The $(k+2)^{th}$ element indicates Y_i, the $(k+2)^{th}\theta$. From (17) 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:

$$P_{\beta,\gamma}(x_{obs,i}, Y_i = y^{(\ell)} = \int_{x_{mis,i}} \int_{\theta_i} P_{\beta,\gamma}(x_{obs,i}, x_{mis,i}, Y_i = y^{(\ell)}, \theta_i) d\theta_i dx_{mis,i} =$$

$$\int_{\theta_i} P_{\beta_{(\ell)}}(x_{obs,i} \mid Y_i = y^{(\ell)}, \theta_i) P_y(\theta_i \mid Y_i = y^{(\ell)}) \mathrm{P}(Y_i = y^{(\ell)}) d\theta_i =$$

$$\int_{\theta_i} P_{\beta_{(\ell)}}(x_{obs,i} \mid \theta_i) g_{y_\ell}(\theta_i) . \pi_\ell d\theta_i =$$

$$\prod_{\ell=1}^{L} \pi_\ell^{y_{i\ell}} . \prod_{\ell=1}^{L} \left\{ \int_{\theta_i} P_{\beta_{(\ell)}}(x_{obs,i} \mid \theta_i) g_{y_\ell}(\theta_i) d\theta_i \right\}^{y_{i\ell}} . \tag{18}$$

The likelihood of the total sample is given by:

$$L(\beta, y, \pi; x_{obs}, y) = \prod_{i=1}^{n} P_{\beta,\gamma}(x_{obs,i}, Y_i = y^{(\ell)}) = \tag{19}$$

$$\prod_{i=1}^{n}\prod_{\ell=1}^{L}\pi_{\ell}^{y_{i\ell}} . \prod_{i=1}^{n}\prod_{\ell=1}^{L}\left\{\int_{\theta_i} P_{\beta_{(\ell)}}(x_{obs,i} \mid \theta_i).g_{y_\ell}(\theta_i)d\theta_i\right\}^{y_{i\ell}} .$$

From (19) it is seen that the likelihood function consist i. of a term that depends only on the proportions π_ℓ of the subpopulations in the total population, and ii. a term which is a product of L ordinary marginal likelihood functions. This is because there is always exactly one ℓ for which $y_{i\ell} = 1$, with the understanding that they not all contain the same item parameters. Standard maximum likelihood estimates $\hat{\pi}_\ell, \ell = 1,...,L$ of the proportions can be obtained from the first part. Maximizing the second term with respect to $y_\ell, \ell = 1,...,L$ and β 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 m_i establishes becomes:

$$\left.\begin{array}{l} U_{obs,i} = X_{obs,i} \\ U_{mis,i} = (X_{mis,i} Y_i, \theta_i) \end{array}\right\}, (i = 1,...,n) \tag{20}$$

The design distribution is given by:

$$\mathrm{P}(M_i = (r_\ell, 0, 0) \mid Y_i = y^{(\ell)}) = \varphi_\ell, \ (\ell = 1,...,L) . \tag{21}$$

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. The next example will illustrate this.

Example 2.

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 $\theta \sim \mathrm{N}(-1,1)$ and $\theta \sim \mathrm{N}(+1,1)$. 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 targeted 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 the third column of Table 2.

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 $(\mathrm{E}(\theta) = -1.0)$ are overestimated and are underestimated in the more able group $(\mathrm{E}(\theta) = 1.0)$. If we do not ignore the design variable and estimate with two marginal distributions (19) we get the results in the fourth column of Table 2, which are seen to be free from systematic bias. It is noted that the results of these two calibrations are comparable by fixing both scales by $\sum_i \beta_i = 0$

Table 2. Input and estimated difficulty parameters Rasch model

item	β (input)	β (se); one marginal	β (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		$\hat{\mu} = 0.047$	$\hat{\mu}_1 = -0.986$ $\hat{\mu}_2 = 1.097$
sd		$\hat{\sigma} = 1.293$	$\hat{\sigma}_1 = 0.954$ $\hat{\sigma}_2 = 1.129$

In the TTMP situation the background variable is used as a stratification variable: from every subpopulation $\ell, \ell = 1,...,L$, we have a random sample from $g_{\gamma_\ell}(\theta)$ with n_ℓ the number of observations in subpopulation ℓ and $\Sigma_{\ell=1}^{L} n_\ell = n$ 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 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 (16) can only completely be estimated provided the population proportions are known and that we have samples

from every subpopulation, $n_\ell > 0, \ell = 1,...,L$. Otherwise we are not able to estimate all subpopulation parameters $\gamma_\ell, \ell = 1,...,L$. Another difference from the TTOP situation is that in TTMP 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 observed response vector $x_{obs,i}$ and Y_i as in the TTOP case (18), but the conditional distribution of X_i given $Y_i = y^{(\ell)}$. The design distribution is given by:

$$\mathrm{P}(M_i = m_{i_\ell} = (r_\ell, 0)) = 1 \text{ if } Y_i = y^{(\ell)}. \tag{22}$$

Compared to (17), the TTOP case has one element less, because is not random. Because of (22) the conditional distribution of a response vector given $Y_i = y^{(\ell)}$ is the same as the conditional distribution of given the design variable. For a randomly sampled student from subpopulation we have:

$$P_{\beta_{(\ell)},\gamma_\ell}(x_{obs,i_\ell} \mid m_{i_\ell}) = P_{\beta_{(\ell)},\gamma_\ell}(x_{obs,i_l} \mid Y_i = y^{(\ell)}) =$$

$$\int_{x_{mis,i\ell}} \int_{\theta_{i_\ell}} P_{\beta_{(\ell)},\gamma_\ell}(x_{obs,i_\ell}, x_{mis,i_\ell}, \theta_{i_\ell}) \mid Y_i = y^{(\ell)}).P_{\gamma_\ell}(\theta_{i_\ell} \mid Y_i = y^{(\ell)}) d\theta_{i_\ell} dx_{mis,i} =$$

$$\int_{\theta_{i_\ell}} P_{\beta_{(\ell)}}(x_{obs,i_\ell} \mid \theta_{i_\ell}).g_{\gamma_\ell}(\theta_{i_\ell}) d\theta_{i_\ell} .$$

And for the total sample we have the likelihood

$$\prod_{\ell=1}^{L} \prod_{i_\ell=1}^{n_\ell} \int_{\theta_{i_\ell}} P_{\beta_{(\ell)}}(x_{obs,i_\ell} \mid \theta_{i_\ell}).g_{\gamma_\ell}(\theta_{i_\ell}) d\theta_{i_\ell} . \tag{23}$$

As before the parameters β and $\gamma_\ell, \ell = 1,...,L$, provided $n_\ell > 0$, can be estimated from (23). 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

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

then we are in fact in the TTOP situation for which it was shown ((20) and (21)) that by Y ignoring the MAR condition for ignorability is not fulfilled.

Summarizing we can say that in MML item calibration in complete testing designs is justified 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 Y must be used. Not using 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 Y, and to distinguish more subgroups in the samples, the awareness of the possible problems is not general and 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 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 that 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 M at the observed pattern of missing data m and assume that the values u_{obs} are realizations of the marginal distribution of U_{obs} (7):

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

Remember (8) that the correct distribution of the realizations u_{obs},

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

the conditional distribution of U_{obs} given $M = m$, is not used in the analysis, but only the marginal distribution of the observed responses. Note that in the CML case, the design variable M_i and the item indicator variable 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

variable in CML estimation is only possible if for an individual observed response vector $X_{obs,i}$ there exists a sufficient statistic $S_{obs,i} = S(X_{obs,i})$ for θ_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

$$S_{obs,i} = \sum_{j \in obs} X_{ij},$$

is not only not sufficient for θ_i in the marginal distribution of the observations $X_{obs,i}$, but also not sufficient in the distribution of all observed data $(X_{obs,i}, R_i)$. $S_{obs,i}$ is only sufficient in the conditional distribution of the responses given the item indicator variable R_i. An example will make this clear. Assume we have 3 items following the Rasch model with parameters $\in_i = \exp(-\beta_i), i = 1,2,3$ and a random item indicator variable with two possible outcomes $(0 < \varphi < 1)$:

$\mathrm{P}(R_i = r_1 = (1,1,0)) = \varphi$, and $\mathrm{P}(R_i = r_2 = (1,0,1)) = 1 - \varphi$.

In Table 3 the relevant probabilities for all outcomes with $S_{obs} = 1$, with $\exp(\theta) = \xi$ are given.

Table 3. Probabilities for all outcomes with $S_{obs} = 1$

x_{obs}, r	$p(x_{obs}, r)$ (i)	$p(x_{obs} \mid r_1)$ (ii)	$p(x_{obs} \mid r_2)$ (iii)
10,110	$\frac{\varphi.\xi \in_1}{(1+\xi \in_1)(1+\xi \in_2)}$	$\frac{\xi \in_1}{(1+\xi \in_1)(1+\xi \in_2)}$	0
01,110	$\frac{\varphi.\xi \in_2}{(1+\xi \in_1)(1+\xi \in_2)}$	$\frac{\xi \in_2}{(1+\xi \in_1)(1+\xi \in_2)}$	0
10,101	$\frac{(1-\varphi).\xi \in_1}{(1+\xi \in_1)(1+\xi \in_3)}$	0	$\frac{\xi \in_1}{(1+\xi \in_1)(1+\xi \in_3)}$
01,101	$\frac{(1-\varphi).\xi \in_3}{(1+\xi \in_1)(1+\xi \in_3)}$	0	$\frac{\xi \in_3}{(1+\xi \in_1)(1+\xi \in_3)}$
1	$\frac{\varphi.\xi(\in_1 + \in_2)}{(1+\xi \in_1)(1+\xi \in_2)} + \frac{(1-\varphi.\xi(\in_1 + \in_3)}{(1+\xi \in_1)(1+\xi \in_3)}$	$\frac{\xi(\in_1 + \in_2)}{(1+\xi \in_1)(1+\xi \in_2)}$	$\frac{\xi(\in_1 + \in_3)}{(1+\xi \in_1)(1+\xi \in_3)}$
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 X_{obs} and R, that is, dividing in Table 3 the terms in the upper part of column (i) 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 X_{obs} given R, S_{obs} is sufficient for ξ. Divide the upper part terms in column (ii) and (iii) in Table 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

$$P_{\theta,\beta,\varphi}(x_{obs}, r) = \prod_i P_{\theta_i,\beta,\varphi}(x_{obs,i} \mid r_i).P_{\varphi}(r_i)\,. \tag{24}$$

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 (24) as:

$$P_{\theta,\beta,\varphi}(x_{obs}, r) = \prod_{t=1}^{T}\prod_{i-1}^{n_t} P_{\theta_i,\beta_{(t)},\varphi}(x_{obs,i} \mid r_t).P_{\varphi}(r_t)\,. \tag{25}$$

We see in (25) that we have in fact the product of T complete data likelihoods. For every t the first factor in the right-hand side of (25) can, as in complete data CML (see (3)), be written as

$$\prod_{i=1}^{n_t} P_{\theta_i,\beta_{(t)},\varphi}(x_{obs}, r_t) = \prod_{i=1}^{n_t} P_{\beta_{(t)}}(x_{obs,i} \mid s_{obs,i}, r_t).P_{\theta_i,\beta_{(t)},\varphi}(s_{obs,i}, r_t)\,. \tag{26}$$

And the first factor in the right-hand side of (26) is again free of any incidental parameters, and

$$L_c = \prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\beta_{(t)}}(x_{obs,i} \mid s_{obs,i}, r_t) \tag{27}$$

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 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 (27) depends of course, as in the complete data case, on the properties of the part of the total likelihood (25) we neglect in that case. The neglected part in CML estimation in incomplete designs is (combining (25), (26) and (27))

$$\prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\theta_i,\beta_{(t)},\varphi}(s_{obs,i}, r_t) = \prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\theta_i,\beta_{(t)},\varphi}(s_{obs,i} \mid r_t).\prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\varphi}(r_t)\,. \tag{28}$$

In (28), 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 (28) for the three considered design types next.

CML in random incomplete designs. In random incomplete designs the design distribution is given by (12). Considering the first factor of the part of the likelihood we neglect in CML (28), we see this factor consists of the product of T complete data distributions of the sufficient statistics s_{obs} , which can be neglected. From the design distribution (12) it is easily seen that the second part of (28), $P_{\varphi}(r_t)$, does not depend on the item parameters at all. As a consequence, (28) 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 (28) 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 (14) we can write the second part as:

$$\prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\varphi}(R_i = r_t) = \prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\varphi}(R_i = r_t \mid x_{obs,i}).P_{\beta_{(obs)},\theta_i}(x_{obs,i})\,. \tag{29}$$

We see that (29) is for every t directly dependent of the item parameters of the items used for establishing the design. This means that (28), cannot be neglected in CML estimation. So CML estimation is in this situation not justified because it implies that not all random variations in the data relevant for estimating the item parameters are considered in the conditional likelihood. Applying CML estimation in these designs, which is possible by running standard computer programs for CML, gives incorrect estimates of the item parameters. An example will illustrate this.

Example 1 (continued).

The items and the design used are given in example 1. 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 the third column in Table 4. In the fourth column the results are given of the item parameter estimates in the two stage testing design.

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.

Table 4. CML estimates and standard errors in a two stage testing design

item	β (input)	β (se) complete	β (se) multistage	β (se) 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)

The last column of Table 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 (26) 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:

$$L_c = \prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\beta^2_{(t)}}(x_{obs2,i} \mid s_{obs2,i}, r_t).$$

And the distribution of the design which is neglected in the estimation is given by

$$\prod_{t=1}^{T}\prod_{i=1}^{n_t} P_{\varphi}(R_i = r_t \mid x_{obs1,i}).P_{\beta_{obs1,i}\theta_i}(x_{obs1,i}).$$

does not depend on the parameters $\beta^{(2)}$, which are estimated.

Following the procedure given in Example 1 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 Y determines the design. The design distribution is given by (14). Before we made the distinction between the two sampling roles Y can play in the design and using or not using 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 X_i and of Y_i of student i is given by

$$P_{\theta_i,\beta,\pi_\ell}(x_i, Y_i = y^{(\ell)}) = P_{\theta_i,\beta}(x_i \mid Y_i = y^{(\ell)}).\mathrm{P}_{\pi_\ell}(Y_i = y^{(\ell)}) .$$

Conditioning on the sufficient statistic S_i gives:

$$\begin{aligned} P_{\theta_i,\beta,\pi_\ell}(x_i, Y_i = y^{(\ell)}) &= P_{\theta_i,\beta}(x_i \mid s_i, Y_i = y^{(\ell)}).P_{\theta_i,\beta}(s_i \mid Y_i = y^{(\ell)}).\mathrm{P}_{\pi_\ell}(Y_i = y^{(\ell)}) \\ &= P_{\beta}(x_i \mid s_i).P_{\theta_i,\beta}(s_i \mid Y_i = y^{(\ell)}).\mathrm{P}_{\pi_\ell}(Y_i = y^{(\ell)}) . \end{aligned} \tag{30}$$

In (30) $Y_i = y^{(\ell)}$ cancels in $P_{\beta}(x_i \mid s_i)$ because given θ_i the item responses are not dependent of any other characteristic of the students (local independence). The complete likelihood of the sample is given by:

$$\prod_i P_{\beta}(x_i \mid s_i).\prod_i P_{\theta_i,\beta}(s_i \mid Y_i = y^{(\ell)}).\mathrm{P}_{\pi_\ell}(Y_i = y^{(\ell)}) . \tag{31}$$

From (31), 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 β. 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 $P_{\theta_i,\beta}(x_i \mid Y_i = y^{(\ell)})$ with as a consequence that $P_{\pi_\ell}(Y_i = y^{(\ell)})$ cancels in (30) and (31). 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 r_i as we distinguish values of the background variable Y_i. If we rewrite the total likelihood as before ((25), (27) and (28)) we see that the conditional likelihood to be maximized is:

$$\prod_{\ell=1}^{L}\prod_{i=1}^{n_\ell} P_{\beta_\ell}(x_{obs,i} \mid s_{obs,i}, r_\ell, Y_i = y_{(\ell)}), \tag{32}$$

and the neglected part becomes

$$\prod_{\ell=1}^{L}\prod_{i=1}^{n_\ell} P_{\theta_i,\beta_{(\ell)},\varphi,\pi}(s_{obs,i}, r_\ell, Y_i = y^{(\ell)}) = \tag{33}$$

$$\prod_{\ell=1}^{L}\prod_{i=1}^{n_\ell} P_{\theta_i,\beta_{(\ell)},\varphi,\pi}(s_{obs,i} \mid r_\ell, Y_i = y^{(\ell)}) = \prod_{\ell=1}^{L}\prod_{i=1}^{n_\ell} P_{\varphi,\pi}(r_\ell, Y_i = y^{(\ell)}).$$

From the design distribution (14) it is seen that the second part of the right hand side of (33) is independent of the item parameters which are to be estimated. So CML estimation, on the basis of the conditional likelihood (32), is justified in targeted testing.

Example 2 (continued)

If we estimate the item parameters of example 2 with CML, we see in results of Table 5. that targeted testing does not cause any systematic errors in the item parameter estimates.

Table 5. Input β and estimated $\hat{\beta}$ difficulty parameters Rasch model

item	β (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)

Conclusion

In this study it has been shown for the three most common stochastic incomplete design types under which conditions item calibration is possible. It was seen that in MML estimation Rubins ignorability principle 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.

Summarizing 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 stage 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 in standard computer algorithms for MML assuming 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. In this study some small examples were given to show the miss-behaviour of some procedures. How much impact this has in practical situations in which we have more items and with other distributions of the item parameters, is worthwhile exploring.

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 study all results are for convenience illustrated by the simple one-parameter logistic model for dichotomously scored items. But the results also apply, whenever CML or MML is applicable, for models for polytomously 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. (2000). On the loss of information in conditional maximum likelihood estimation. *Psychometrika*, 65, 337-362.

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.

Hanson, B.A. & Béguin, A.A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent calibration in the common-item equating design. *Applied Psychological Measurement*, 26, 3-24

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.

Mislevy, R.J. & Wu, P-K (1996). Inferring examinee ability when some item responses are missing. *Research Report RR-96-30-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 logististic model (OPLM).* [Computer software]. Arnhem: Cito.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG [Computer Software]*. Chicago: Scientific Software International.

(Manuscript received: 18 July 2008; accepted: 22 March 2010)