Computerized adaptive testing for the random weights linear logistic test model

Marjolein Crabbe, Martina Vandebroek

Computerized adaptive testing for the random weights linear logistic test model

Marjolein Crabbe ^{a1}

Email: marjolein.crabbe@kuleuven.be

Tel: +32 16 326961

Fax: +32 16 326624

Martina Vandebroek ^{a,b}

Email: martina.vandebroek@kuleuven.be

^a Faculty of Business and Economics, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

^b Leuven Statistics Research Centre, KU Leuven, Celestijnenlaan 200B, 3001 Leuven, Belgium

Research funded by ZKC1090 / DBOF/08/014 - DBOF project of the KU Leuven

¹Corresponding author

Abstract

This paper discusses four item selection rules to design efficient individualized tests for the random weights linear logistic test model: minimum posterior weighted ($\mathcal{D}_B$) and minimum expected posterior weighted ($E\mathcal{D}_B$) $\mathcal{D}$ -error, maximum expected Kullback-Leibler divergence between subsequent posteriors (KLP) and maximum mutual information (MUI). The random weights linear logistic test model decomposes test items into a set of subtasks or cognitive features and assumes individual-specific effects of the features on the difficulty of the items. In contrast to a single ability score, the individual effects provide a more profound profile of a test taker’s proficiency, giving one’s strengths and weaknesses with respect to the item features. Simulations show how the design efficiency of the different criteria appears to be equivalent. However, KLP and MUI are given preference over $\mathcal{D}_B$ and $E\mathcal{D}_B$ due to their lower complexity, highly reducing the computational intensity.

Keywords: random weights linear logistic test model, cognitive diagnosis, computerized adaptive testing, $\mathcal{D}$ -efficiency, Kullback-Leibler divergence, mutual information

1 Introduction

The linear logistic test model (LLTM) achieved great popularity and value in item response theory (IRT) for item complexity modeling. Since its introduction by Fischer in 1973, the LLTM has been applied in various educational and psychological settings to analyze the cognitive structure of test items, more specifically the decomposition of test items into a specific set of item stimulus features or subtasks and the influence of these features on the difficulty of the items (Bouwmeester et al., 2011; Chen et al., 2011; Embretson and Daniel, 2008; Freund et al., 2008; Holling et al., 2008; Holling et al., 2009; Hornke and Habon, 1986; Medina-Diaz, 1993; Poinstingl, 2009; Sonnleitner, 2008; Whitely and Schneider, 1981; Zeuch et al., 2011). But despite the LLTM’s ease of interpretation and its many successful and instructive applications, one might question the assumption of fixed feature effects. Divergent abilities might cause the presence of specific subtasks in an item to affect people (and their probability to correctly solve

the item) in a different way. Because of these possible differences in attribute weights between individuals, Rijmen and De Boeck (2002) extended the LLTM to the random weights linear logistic test model (RWLLTM) which allows individual feature effects and therefore accounts for potential heterogeneity in the aptitude of test takers.

By modeling individual feature effects on item difficulty, the RWLLTM measures in essence test takers' strengths and weaknesses with respect to the subtasks involved in the test items or, alternatively stated, with respect to different dimensions of ability. In this light the RWLLTM can be considered a cognitive diagnosis model. Cognitive diagnosis is a topical subject in the IRT literature and stems from an increasing interest in a more profound profile of individuals' proficiency. In addition to merely auditing and ranking one's ability with a single test score, cognitive diagnosis gives insight and feedback on the cognitive skills the examinees master and the ones they are struggling with. With this additional information one is able to tailor instruction and improve individual learning. Other examples of recently developed cognitive diagnosis models are the DINA, the NIDA and the fusion model (de la Torre, 2009; de la Torre and Douglas, 2008; Henson and Douglas, 2005; McGlohen and Chang, 2008). Moreover, multidimensional IRT (MIRT) models, such as the multidimensional two- and three-parameter logistic model (Mulder and van der Linden, 2009; Segall, 2010; Wang and Chang, 2011; Wang et al., 2011; Yao and Boughton, 2007), also analyze multiple dimensions in the ability space.

The advantage of the RWLLTM over MIRT models is however that ability estimation in the former does not rely on item calibration. Like other diagnostic models, the RWLLTM characterizes test items by the presence or absence of pre-defined features, which unambiguously links the coefficients in the model to one of these specific subtasks. In MIRT models on the other hand, one must assume that the pre-calibration of the item pool yields discrimination scores that represent the abilities of interest. MIRT may well improve the accuracy of trait estimation in comparison to unidimensional models, one is however not sure whether the latent traits measured are indeed the ones one wants to measure (Li and Schafer, 2005; Wang et al., 2011). Moreover, with the RWLLTM one attains a continuous ability parameter for each subfeature. Continuous estimates, which are also obtained with MIRT models, provide more information

about the skills under diagnosis than the simple dichotomous mastery / non-mastery scores from cognitive diagnosis models (Wang et al., 2011). Combining the benefits of both cognitive diagnosis and MIRT models, the RWLLTM has great potential for test analysis.

Therefore, this research discusses the efficient design of tests to accurately estimate the RWLLTM. Four different item selection rules are compared to construct individualized tests by means of computerized adaptive testing (CAT). The tests are sequentially designed for each person separately, using prior information on the parameters and one's responses to previous test items. As such, the tests are tailored to the specific abilities of an individual, resulting in higher quality test data and in turn in more accurate and more valuable diagnostic profiles. The first two design algorithms apply the well known $\mathcal{D}$ -efficiency criterion and maximize the determinant of the model's Fisher information matrix. Note that one weighs the criterion over Bayesian updates of the posterior distribution of the random effects (van der Linden, 1998; van der Linden and Pashley, 2010), whereas the other considers expected updates of posteriors to weigh the $\mathcal{D}$ -criterion (van der Linden, 1998; van der Linden and Pashley, 2010). In contrast, the third and the fourth item selection rule are based on Kullback-Leibler divergence between subsequent posterior distributions (Mulder and van der Linden, 2010; Wang and Chang, 2011). Markov Chain Monte Carlo (MCMC) estimation is used to estimate the RWLLTM.

The criteria above are not new in adaptive IRT design and have already been applied to a couple of item response models. Van der Linden (1998) and van der Linden and Pashley (2010) for example compare the design criteria based on $\mathcal{D}$ -efficiency for respectively the two- and the three-parameter logistic model. Both researches show the benefit of weighing the criterion over expected updates of the posterior distribution. Wang and Chang (2011) on the other hand compare both Fisher information and Kullback-Leibler information item selection criteria for the multidimensional three-parameter logistic model. Note however that instead of weighing $\mathcal{D}$ -efficiency over posteriors, the Fisher information is evaluated at posterior modes in this paper yielding only locally efficient designs. They found improved estimation accuracy with the Kullback-Leibler criteria. As the RWLLTM is both conceptually and analytically very different from the models in previous researches and, moreover, shows great advantages over the

increasingly popular MIRT and cognitive diagnosis models, this study provides an instructive extension in the research on individualized test design.

The remainder of this paper first introduces the random weights linear logistic test model and the four item selection rules. The third section includes a simulation study on the estimation performance of the MCMC algorithm used to estimate the models. The main study comparing the design criteria in constructing efficient individualized designs for the RWLLTM is discussed in section four. Results with respect to estimation accuracy, item exposure and item overlap are given. The final section discusses the main findings.

2 Method

2.1 The random weights linear logistic test model

Fischer's (1973) linear logistic test model (LLTM) assumes that test items can be decomposed into a set of subtasks, also denoted as cognitive operations, item rules, attributes or features. For each item in a test it can then be indicated whether or not a specific subtask must be completed in order to get the right answer. Consequently, item difficulty is defined as a weighted sum of the required features with weights expressing the relative effect of a subtask on the difficulty of the items. In the LLTM, the probability that person i solves item j correctly is modeled as (Fischer, 1973)

$$P(Y_{ij} = 1 | \mathbf{x}_j; \theta_i, \boldsymbol{\beta}) = \frac{\exp(\theta_i + \sum_m \beta_m x_{jm})}{1 + \exp(\theta_i + \sum_m \beta_m x_{jm})} \quad (1)$$

$$= \frac{\exp(\theta_i + \mathbf{x}'_j \boldsymbol{\beta})}{1 + \exp(\theta_i + \mathbf{x}'_j \boldsymbol{\beta})}, \quad (2)$$

with θ_i the ability parameter for person i measuring overall proficiency, x_{jm} the score of item j on item feature m , which is 1 if subtask m is required to solve item j correctly and 0 otherwise, and β_m the weight corresponding to feature m 's relative effect on item difficulty. Note that the Rasch model essentially is a special case of the LLTM, assuming each item in a test represents a subtask. Instead of difficulty weights for the item features, item difficulty parameters are

estimated for each item separately. Once the cognitive structure of test items is determined by the LLTM, (infinitely many) items with a specific difficulty level can be constructed.

Recently however, some alternative applications of the LLTM have also been discussed in the test theory literature (Kubinger, 2008; Kubinger, 2009). The model appears for instance to be a proper means for examining learning and fatigue effects, concerning the position of items in a test (Hahne, 2008; Hohensinn et al., 2008), age effects (Kubinger et al., 2011) and the effect of applying different response formats (Hohensinn and Kubinger, 2009; Hohensinn and Kubinger, 2011).

The fixed effects assumption of specific features on item difficulty in (2) might however not always be realistic. A clear example are mathematical questions for which the problem is described in a short text and thus requiring both mathematical and verbal skills to solve them (Rijmen and De Boeck, 2002). As not all persons have the same analytical or linguistic competence, it is no surprise that heterogeneity in feature effects may arise. To account for this, the random weights linear logistic test model (RWLLTM) extends the LLTM and allows person-specific effects of the item features on the probability of success. The random effects measure the differences in subtask difficulty between individuals and therefore in individuals' abilities. The RWLLTM thus serves as a cognitive diagnosis model, providing for each individual a more informative diagnostic profile with respect to the skills of interest.

Conditional on $\mathbf{b}_i$, the feature effects for individual i , the RWLLTM probability that person i passes item j then is (Rijmen and De Boeck, 2002)

$$P(Y_{ij} = 1 | \mathbf{x}_j; \theta_i, \boldsymbol{\beta}, \mathbf{b}_i) = \frac{\exp(\theta_i + \mathbf{x}'_j \boldsymbol{\beta} + \mathbf{z}'_j \mathbf{b}_i)}{1 + \exp(\theta_i + \mathbf{x}'_j \boldsymbol{\beta} + \mathbf{z}'_j \mathbf{b}_i)}, \quad (3)$$

with $\mathbf{z}_j$ a subset of $\mathbf{x}_j$ including the item features with a random instead of a fixed effect. The coefficient θ_i measures ability not specified by the item features (Henson and Douglas, 2005; McGlohen and Chang, 2008). Further, the model assumes a heterogeneity distribution, in most cases a multivariate normal distribution, over the individual-specific coefficients

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}). \quad (4)$$

By assuming a zero mean in the random effects distribution, $\boldsymbol{\beta}$ contains the weights of the features with a fixed effect and the mean effects of features with a random effect. Moreover, as the examinees' overall ability θ_i is also random, it can be incorporated in the vector of individual effects $\mathbf{b}_i$. Taking $x_{j0} = z_{j0} = 1$ for all items and $b_{i0} \equiv \theta_i$, the model in (3) becomes

$$P(Y_{ij} = 1 | \mathbf{x}_j; \boldsymbol{\beta}, \mathbf{b}_i) = \frac{\exp(\mathbf{x}'_j \boldsymbol{\beta} + \mathbf{z}'_j \mathbf{b}_i)}{1 + \exp(\mathbf{x}'_j \boldsymbol{\beta} + \mathbf{z}'_j \mathbf{b}_i)}. \quad (5)$$

In the remainder of the text we will assume $\mathbf{x}_j$ is p -dimensional and $\mathbf{z}_j$ a $q \times 1$ subset of it.

Finally, the likelihood of the RWLLTM for a sample of N test takers and a test including J items is

$$L(\boldsymbol{\beta}, \mathbf{D} | \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^N \int L(\boldsymbol{\beta}, \mathbf{b}_i | \mathbf{y}_i, \mathbf{X}) \phi(\mathbf{b}_i | \mathbf{0}, \mathbf{D}) d\mathbf{b}_i \quad (6)$$

$$= \prod_{i=1}^N \int \prod_{j=1}^J \frac{\exp[y_{ij}(\mathbf{x}'_j \boldsymbol{\beta} + \mathbf{z}'_j \mathbf{b}_i)]}{1 + \exp(\mathbf{x}'_j \boldsymbol{\beta} + \mathbf{z}'_j \mathbf{b}_i)} \phi(\mathbf{b}_i | \mathbf{0}, \mathbf{D}) d\mathbf{b}_i, \quad (7)$$

with $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})'$ the response vector of person i and y_{ij} 1 if person i correctly solves item j and 0 otherwise, $\mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_J)'$ the test design and ϕ the normal density. MCMC estimation, more specifically Gibbs sampling, is used to estimate the RWLLTM. Details on the estimation procedure are given in Appendix A.

2.2 Computerized adaptive testing for the RWLLTM

As an important objective of modeling a RWLLTM is to reveal persons' strengths and weaknesses with respect to the cognitive test features under study, accurate estimation of the random effects $\mathbf{b}_i$ is necessary. Therefore, instead of designing a fixed test for all examinees, it is more sensible and efficient to construct individualized tests customized to the specific abilities of the examinees, improving the quality of the test data. Though the sequential design of tests is a very topical subject, it made its appearance in item response theory already decades ago (see for instance Berger, 1992). The next sections describe four different item selection rules to construct efficient individualized test designs for the RWLLTM. The methods will be compared

in a simulation study with respect to their design efficiency, item exposure and item overlap.

2.2.1 Minimum posterior weighted $\mathcal{D}$ -error

The $\mathcal{D}$ -optimality criterion (Atkinson et al., 2007) is a well known and widely used measure to construct efficient designs. In a multidimensional setting this criterion equals the determinant of the model's Fisher information matrix, which is the negative expectation of the second partial derivative of the log-likelihood function. The rationale behind this criterion is that the determinant of the Fisher information matrix is inversely proportional to the volume of the confidence ellipsoid around the maximum likelihood parameter estimates. Therefore, an efficient test design is a design that maximizes the $\mathcal{D}$ -optimality criterion. Note that here, as Bayesian inference is assumed, the logarithm of the posterior instead of the log-likelihood will be incorporated in the criterion (Mulder and van der Linden, 2009; Segall, 2010; Yu et al., 2011) yielding a Bayesian Fisher information matrix (BFIM).

The information matrix for a test design $\mathbf{X}$ with respect to the random coefficients $\mathbf{b}$ of an individual is then given by

$$\mathbf{I}_{BFIM}(\mathbf{b}, \mathbf{X}) = -E \left[\frac{\partial^2 \log[L(\boldsymbol{\beta}, \mathbf{b} | \mathbf{y}, \mathbf{X}) f_0(\mathbf{b})]}{\partial \mathbf{b} \partial \mathbf{b}'} \right], \quad (8)$$

with $f_0(\mathbf{b})$ a prior distribution for the random feature effects. Note that the index i is dropped to simplify notation. Assuming prior values $\boldsymbol{\beta}_0$ for the mean effects and $f_0(\mathbf{b}) \equiv \phi(\mathbf{b} | \mathbf{0}, \mathbf{D}_0)$, the normal density with variance covariance matrix $\mathbf{D}_0$, one can show that

$$\mathbf{I}_{BFIM}(\mathbf{b}, \mathbf{X}) = \sum_{j=1}^J \frac{\exp(\mathbf{x}'_j \boldsymbol{\beta}_0 + \mathbf{z}'_j \mathbf{b})}{[1 + \exp(\mathbf{x}'_j \boldsymbol{\beta}_0 + \mathbf{z}'_j \mathbf{b})]^2} \mathbf{z}_j \mathbf{z}'_j + \mathbf{D}_0^{-1}. \quad (9)$$

Equivalent to maximizing the determinant of $\mathbf{I}_{BFIM}$, we will minimize the inverse, denoted as the $\mathcal{D}$ -error,

$$\mathcal{D} = \det[\mathbf{I}_{BFIM}(\mathbf{b}, \mathbf{X})]^{-1/q}. \quad (10)$$

Moreover, instead of fixing the random effects $\mathbf{b}$ in the criterion to some specific prior values or estimates and merely obtaining locally efficient designs, the $\mathcal{D}$ -error is averaged over a prior distribution of the random effects which is updated each time an item response is observed.

As no answers are available at the beginning of the test, the Bayesian $\mathcal{D}$ -error is minimized over all possible test items with respect to a preset prior distribution $f(\mathbf{b})$ of the random effects to obtain the first item

$$\mathcal{D}_B = \int \det[\mathbf{I}_{BFIM}(\mathbf{b}, \mathbf{x}_1)]^{-1/q} f(\mathbf{b}) d\mathbf{b}. \quad (11)$$

For this and all subsequent design criteria, we set the initial prior $f(\mathbf{b})$ to $\phi(\mathbf{b}|\mathbf{0}, \mathbf{D}_0)$. Note that although here this “design” prior equals the “inference” prior $f_0(\mathbf{b})$ in (9), this is not necessary. After $k - 1$ test items have been solved by the examinee, the prior information on $\mathbf{b}$ can be updated with the observed response data in the following way

$$f(\mathbf{b}|\mathbf{y}_{k-1}) = \frac{L(\beta_0, \mathbf{b}|\mathbf{y}_{k-1}, \mathbf{X}_{k-1}) \phi(\mathbf{b}|\mathbf{0}, \mathbf{D}_0)}{\int L(\beta_0, \mathbf{b}|\mathbf{y}_{k-1}, \mathbf{X}_{k-1}) \phi(\mathbf{b}|\mathbf{0}, \mathbf{D}_0) d\mathbf{b}}, \quad (12)$$

with $\mathbf{y}_{k-1} = (y_1, \dots, y_{k-1})'$. The k th item in the test is now obtained by minimizing the Bayesian $\mathcal{D}$ -error (for the design $\mathbf{X}_k$ including the $k - 1$ previously administered test items and the current candidate for item k), weighted over this posterior distribution:

$$\mathcal{D}_B = \int \det[\mathbf{I}_{BFIM}(\mathbf{b}, \mathbf{X}_k)]^{-1/q} f(\mathbf{b}|\mathbf{y}_{k-1}) d\mathbf{b}. \quad (13)$$

2.2.2 Minimum expected posterior weighted $\mathcal{D}$ -error

The design algorithm above updates the posterior distribution of the random effects with the observed responses from previous test items. The current criterion on the other hand updates posteriors up until the candidate item, averaging over all possible responses to that item. As only dichotomous test items are assumed here, the answer is either correct (1) or wrong (0). The k th test item is thus found by minimizing the expected posterior weighted $\mathcal{D}$ -error (van

der Linden, 1998; van der Linden and Pashley, 2010), which is given by

$$E\mathcal{D}_B = \sum_{y_k=0}^1 f(y_k|\mathbf{y}_{k-1}) \int \det[\mathbf{I}_{BFIM}(\mathbf{b}, \mathbf{X}_k)]^{-1/q} f(\mathbf{b}|\mathbf{y}_{k-1}, y_k) d\mathbf{b}, \quad (14)$$

with

$$f(y_k|\mathbf{y}_{k-1}) = \int f(y_k|\mathbf{b}) f(\mathbf{b}|\mathbf{y}_{k-1}) d\mathbf{b} \quad (15)$$

$$f(y_k|\mathbf{b}) = p_k^{y_k} (1 - p_k)^{1-y_k} \quad (16)$$

$$p_k = P(Y_k = 1|\mathbf{x}_k; \boldsymbol{\beta}_0, \mathbf{b}). \quad (17)$$

Van der Linden (1998) and van der Linden and Pashley (2010) denote the weights $f(y_k|\mathbf{y}_{k-1})$, $y_k = 0, 1$ in (14) as the posterior predictive probability function.

2.2.3 Maximum expected Kullback-Leibler divergence between subsequent posteriors

Chang and Ying (1996) introduced a totally different approach for optimal item selection in individualized test design using Kullback-Leibler information to distinguish between test items. The Kullback-Leibler divergence between two densities f and g for a continuous variable X equals (Mulder and van der Linden, 2010)

$$KL(f, g) = E_f \left[\log \frac{f(x)}{g(x)} \right] \quad (18)$$

$$= \int f(x) \log \frac{f(x)}{g(x)} dx. \quad (19)$$

Due to the properties that for any two f and g , $KL(f, g) \geq 0$ and that $KL(f, f) = 0$, Kullback-Leibler divergence is commonly interpreted as the distance between two densities. Note however that, in contrast to real distance measures, $KL(f, g) \neq KL(g, f)$ making Kullback-Leibler non-symmetric.

Mulder and van der Linden (2010) extended the ideas of Chang and Ying (1996) and applied Kullback-Leibler divergence to subsequent posteriors of the individual coefficients in an

item response model. They argued that as the subsequent item in a test should extract as much additional information as possible, it should maximize the distance between the current posterior distribution of the individual coefficients and the updated posterior one obtains with the response to the next item. Similar as for the expected posterior weighted $\mathcal{D}$ -error, they average the Kullback-Leibler divergence between posteriors over the possible responses to the candidate item weighing both options (correct or wrong) with the posterior predictive probability function. The k th item is thus found by maximizing the expected Kullback-Leibler distance between subsequent posteriors, i.e.

$$KLP = \sum_{y_k=0}^1 f(y_k|\mathbf{y}_{k-1}) KL[f(\mathbf{b}|\mathbf{y}_{k-1}), f(\mathbf{b}|\mathbf{y}_{k-1}, y_k)]. \quad (20)$$

For practical use, the criterion can be rewritten as (Mulder and van der Linden, 2010)

$$KLP = \sum_{y_k=0}^1 f(y_k|\mathbf{y}_{k-1}) KL[f(\mathbf{b}|\mathbf{y}_{k-1}), f(\mathbf{b}|\mathbf{y}_{k-1}, y_k)] \quad (21)$$

$$= \sum_{y_k=0}^1 f(y_k|\mathbf{y}_{k-1}) \int f(\mathbf{b}|\mathbf{y}_{k-1}) \log \frac{f(\mathbf{b}|\mathbf{y}_{k-1})}{f(\mathbf{b}|\mathbf{y}_{k-1}, y_k)} d\mathbf{b} \quad (22)$$

$$= \sum_{y_k=0}^1 f(y_k|\mathbf{y}_{k-1}) \int \log \frac{f(\mathbf{b}|\mathbf{y}_{k-1})f(y_k|\mathbf{y}_{k-1})}{f(\mathbf{b}, y_k|\mathbf{y}_{k-1})} f(\mathbf{b}|\mathbf{y}_{k-1}) d\mathbf{b} \quad (23)$$

$$= \sum_{y_k=0}^1 f(y_k|\mathbf{y}_{k-1}) \int \log \frac{f(y_k|\mathbf{y}_{k-1})}{f(y_k|\mathbf{b})} f(\mathbf{b}|\mathbf{y}_{k-1}) d\mathbf{b} \quad (24)$$

$$= f(0|\mathbf{y}_{k-1}) \log f(0|\mathbf{y}_{k-1}) + f(1|\mathbf{y}_{k-1}) \log f(1|\mathbf{y}_{k-1}) \\ - f(0|\mathbf{y}_{k-1}) \int \log (1 - p_k) f(\mathbf{b}|\mathbf{y}_{k-1}) d\mathbf{b} - f(1|\mathbf{y}_{k-1}) \int \log p_k f(\mathbf{b}|\mathbf{y}_{k-1}) d\mathbf{b}. \quad (25)$$

2.2.4 Maximum mutual information

Related to Kullback-Leibler divergence is the mutual information between two variables X and Y , defined as (Weissman, 2007)

$$I_M(X, Y) = \int_Y \int_X f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy. \quad (26)$$

It is the Kullback-Leibler distance between the joint distribution of X and Y and their distribution in case of independence and measures how much information the variables capture on one another.

Transforming this into an item selection criterion, Mulder and van der Linden (2010) and Wang and Chang (2011) suggest maximizing the mutual information between the posterior distribution of the random effects and the distribution of the response to the candidate item, given previous responses,

$$MUI = \sum_{y_k=0}^1 \int f(\mathbf{b}, y_k | \mathbf{y}_{k-1}) \log \frac{f(\mathbf{b}, y_k | \mathbf{y}_{k-1})}{f(\mathbf{b} | \mathbf{y}_{k-1}) f(y_k | \mathbf{y}_{k-1})} d\mathbf{b} \quad (27)$$

$$= \sum_{y_k=0}^1 \int f(y_k | \mathbf{b}) f(\mathbf{b} | \mathbf{y}_{k-1}) \log \frac{f(y_k | \mathbf{b}) f(\mathbf{b} | \mathbf{y}_{k-1})}{f(\mathbf{b} | \mathbf{y}_{k-1}) f(y_k | \mathbf{y}_{k-1})} d\mathbf{b} \quad (28)$$

$$= \sum_{y_k=0}^1 \int f(y_k | \mathbf{b}) \log \frac{f(y_k | \mathbf{b})}{f(y_k | \mathbf{y}_{k-1})} f(\mathbf{b} | \mathbf{y}_{k-1}) d\mathbf{b} \quad (29)$$

$$= \int (1 - p_k) \log (1 - p_k) f(\mathbf{b} | \mathbf{y}_{k-1}) d\mathbf{b} + \int p_k \log p_k f(\mathbf{b} | \mathbf{y}_{k-1}) d\mathbf{b} - f(0 | \mathbf{y}_{k-1}) \log f(0 | \mathbf{y}_{k-1}) - f(1 | \mathbf{y}_{k-1}) \log f(1 | \mathbf{y}_{k-1}). \quad (30)$$

The relation of mutual information to Kullback-Leibler distance becomes even more clear from the following calculations, proving that the mutual information criterion is in essence the expected Kullback-Leibler distance between the updated and the current posterior (Mulder and van der Linden, 2010; Wang and Chang, 2011):

$$MUI = \sum_{y_k=0}^1 \int f(\mathbf{b}, y_k | \mathbf{y}_{k-1}) \log \frac{f(\mathbf{b}, y_k | \mathbf{y}_{k-1})}{f(\mathbf{b} | \mathbf{y}_{k-1}) f(y_k | \mathbf{y}_{k-1})} d\mathbf{b} \quad (31)$$

$$= \sum_{y_k=0}^1 \int f(\mathbf{b} | \mathbf{y}_{k-1}, y_k) f(y_k | \mathbf{y}_{k-1}) \log \frac{f(\mathbf{b} | \mathbf{y}_{k-1}, y_k) f(y_k | \mathbf{y}_{k-1})}{f(\mathbf{b} | \mathbf{y}_{k-1}) f(y_k | \mathbf{y}_{k-1})} d\mathbf{b} \quad (32)$$

$$= \sum_{y_k=0}^1 f(y_k | \mathbf{y}_{k-1}) \int f(\mathbf{b} | \mathbf{y}_{k-1}, y_k) \log \frac{f(\mathbf{b} | \mathbf{y}_{k-1}, y_k)}{f(\mathbf{b} | \mathbf{y}_{k-1})} d\mathbf{b} \quad (33)$$

$$= \sum_{y_k=0}^1 f(y_k | \mathbf{y}_{k-1}) KL[f(\mathbf{b} | \mathbf{y}_{k-1}, y_k), f(\mathbf{b} | \mathbf{y}_{k-1})]. \quad (34)$$

Although KLP and MUI are thus very similar (only the order of the posteriors in (20) and (34) is reversed), they are not equal as Kullback-Leibler is a non-symmetrical measure.

Further, Mulder and van der Linden (2010), Wang and Chang (2011) and Weissman (2007) relate mutual information to entropy, a measure of uncertainty and for a continuous variable X defined by

$$H(X) = - \int f(x) \log f(x) dx. \quad (35)$$

They show that MUI is equivalent to

$$MUI = \sum_{y_k=0}^1 f(y_k|\mathbf{y}_{k-1}) \left[\int f(\mathbf{b}|\mathbf{y}_{k-1}, y_k) \log f(\mathbf{b}|\mathbf{y}_{k-1}, y_k) d\mathbf{b} - \right. \quad (36)$$

$$\left. \int f(\mathbf{b}|\mathbf{y}_{k-1}, y_k) \log f(\mathbf{b}|\mathbf{y}_{k-1}) d\mathbf{b} \right] \quad (37)$$

$$= \sum_{y_k=0}^1 f(y_k|\mathbf{y}_{k-1}) [H(\mathbf{b}|\mathbf{y}_{k-1}) - H(\mathbf{b}|\mathbf{y}_{k-1}, y_k)]. \quad (38)$$

The item that maximizes the mutual information criterion is therefore the item that maximizes the expected decrease in uncertainty about $\mathbf{b}$.

Note that the criteria in (25) and (30) only involve posterior weighted response probabilities for the candidate item, whereas the $\mathcal{D}$ -error in the $\mathcal{D}_B$ and ED_B criterion includes also all previously administered items in the test. Clearly, $\mathcal{D}_B$ and ED_B are computationally more intensive than KLP and MUI , which negatively affects their computation time. In practice, the integrals in the optimality criteria are approximated by averages over draws from the distribution at hand. In case of a normal prior at the beginning of the test, draws are easily obtained. Unfortunately, the updated posteriors one gets after each observed response do not have a closed-form expression which complicates sampling. Therefore, importance sampling (Yu et al., 2011) is used to approximate the integrals. Details on this approximation technique are given in Appendix B. Other approaches, such as normal approximation (Mulder and van der Linden, 2010) or Gauss-Hermite quadrature (Wang et al., 2011), might also be applied.

3 Simulation study on MCMC estimation

Before proceeding to the comparison of the different item selection criteria, some results are reported on the performance of the Gibbs sampler in estimating the RWLLTM. Alternative to MCMC, De Boeck et al. (2011) estimate a wide variety of item response models, including the RWLLTM, by maximizing the marginal maximum likelihood (MML) function with Laplace approximation, which is implemented in the `lmer` function of the `lme4` package in R. To demonstrate proper functioning of our Gibbs sampler, its estimation accuracy is compared with that of MML estimation in a simulation study. The study applies a data set on verbal aggression, given in De Boeck et al. (2011) to illustrate the use of the `lmer` function.

The verbal aggression data set consists of the responses of 316 test takers to 24 items on verbal aggression. All items represent a frustrating situation and subjects are asked whether or not they would or would want to react in a verbally aggressive way. There are four different situations, two in which one is self to blame (i.e. “The grocery store closes just as I am about to enter” and “The operator disconnects me when I had used up my last 10 cents for a call”) and two in which someone else is to blame (i.e. “A bus fails to stop for me” and “I miss a train because a clerk gave me faulty information”). Possible verbal aggressive response is either cursing, scolding or shouting and this behavior is presented in either a wanting or a doing mode. A specific item from the questionnaire is for example “A bus fails to stop for me. I want to shout”. This item corresponds to a situation in which someone else is to blame, given with shouting response behavior and a wanting mode. The four different situations, combined with the three behavior types and the two modes yield a test design with 24 items. Test takers could respond with either “yes”, “perhaps” or “no” but responses were dichotomized afterwards to 1 for “yes” and “perhaps” and to 0 for “no”.

To compare the accuracy of the Gibbs sampler and MML estimation from the `lmer` function, two models are considered for the verbal aggression data: a RWLLTM with all feature effects random and a LLTM including only a random intercept, in essence a special case of the former. Note that all items in the test are characterized by four binary features, i.e. one to indicate whether the situation is self-to-blame or other-to-blame, two for representing the behavior type

(cursing, scolding or shouting) and one giving the mode of the item (wanting or doing). In addition, an attribute is included for the random intercept (representing individuals' general verbal aggressiveness) which is set to 1 for all items in the test. Consequently, the mean effects vector β in both the RWLLTM and the LLTM is five-dimensional. Therefore there are also five random coefficients ($q = 5$) in the RWLLTM, whereas in the LLTM only the intercept is random ($q = 1$).

To start, both the RWLLTM and the LLTM were estimated with the `lmer` function for the 316 subjects in the verbal aggression data set. Some merit was found in considering all feature effects random as the AIC value of the RWLLTM and the LLTM respectively equal 7847 and 8250. Their BIC values are 7985 and 8292. In terms of both information criteria, the RWLLTM is thus preferred. This might indicate that by merely assuming fixed effects for the features not all variability in the item responses is captured. Individuals might experience the situations differently or might be heterogeneous in verbal aggressiveness. This is reflected in distinct effects of the item features on their response behavior.

Next, the estimates obtained for the mean effects and the variance covariance matrix were assumed the true coefficients in the population, either denoted as $(\beta_{\text{RWLLTM}}, \mathbf{D}_{\text{RWLLTM}})$ for the RWLLTM or as $(\beta_{\text{LLTM}}, d_{\text{LLTM}})$ for the LLTM, and used to simulate responses for 300, 500, 1000, 1500 and 2000 fictitious subjects for the 24 items in the questionnaire. More specifically, responses were simulated with mean effects β_{RWLLTM} or β_{LLTM} and individual effects $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_{\text{RWLLTM}})$ or $b_i \sim \mathcal{N}(0, d_{\text{LLTM}})$. Note that in case of the LLTM the variance covariance matrix reduces to a single variance as the intercept is the only random coefficient. For each amount of test takers, the simulation of random effects and corresponding responses was repeated 100 times. Finally, for all simulated data sets either a RWLLTM or a LLTM was estimated with both the Gibbs sampler and MML estimation.

Root-mean-squared-errors (RMSE) are used to evaluate and compare the accuracy of both estimation techniques (Chen et al., 2000; Mulder and van der Linden, 2009; Wang and Chang, 2011). They are a measure of total estimation error and are computed separately for all $\mathbf{b}_i$, β

and $\mathbf{D}$. For the individual effects $\mathbf{b}_i$, the RMSE equals

$$\text{RMSE}_{\mathbf{b}} = \sqrt{\frac{1}{N} \sum_{i=1}^N \frac{(\hat{\mathbf{b}}_i - \mathbf{b}_i)'(\hat{\mathbf{b}}_i - \mathbf{b}_i)}{q}}, \quad (39)$$

with $\hat{\mathbf{b}}_i$ and $\mathbf{b}_i$ respectively the estimates and the true (simulated) values for individual i 's random coefficients, N the number of test takers (300, 500, 1000, 1500 or 2000) and q the number of random effects. Next,

$$\text{RMSE}_{\boldsymbol{\beta}} = \sqrt{\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p}}, \quad (40)$$

with $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ respectively the estimated and true mean effects, either $\boldsymbol{\beta}_{\text{RWLLTM}}$ or $\boldsymbol{\beta}_{\text{LLTM}}$, and p the number of attributes.

Finally, to compute the RMSE for $\mathbf{D}$, the variance covariance matrix is transformed into a vector $\mathbf{d}$ containing all unique elements of the matrix. $\text{RMSE}_{\mathbf{D}}$ is given by

$$\text{RMSE}_{\mathbf{D}} = \sqrt{\frac{(\hat{\mathbf{d}} - \mathbf{d})'(\hat{\mathbf{d}} - \mathbf{d})}{q_d}}, \quad (41)$$

with $\hat{\mathbf{d}}$ the estimates and $\mathbf{d}$ the corresponding true coefficients, either the vector transformation of $\mathbf{D}_{\text{RWLLTM}}$ or $\mathbf{D}_{\text{LLTM}}$, and q_d the number of unique elements in the variance covariance matrix, i.e. $q(q+1)/2$.

The RMSEs for $\mathbf{b}_i$, $\boldsymbol{\beta}$ and $\mathbf{D}$ for both MCMC and MML estimation are plotted in Figure 1. Note that the plots display the average estimation errors over the simulation repetitions. The top plots in Figure 1 include the estimation errors for the RWLLTM, results for the LLTM are in the bottom plots.

For the LLTM it can be concluded that the MCMC algorithm is on average as accurate as MML estimation as no significant differences in $\text{RMSE}_{\mathbf{b}}$, $\text{RMSE}_{\boldsymbol{\beta}}$ and $\text{RMSE}_{\mathbf{D}}$ are observed. Also for estimating the mean effects in the RWLLTM, MCMC and MML estimation perform equally well. Yet, with respect to the individual feature effects and the variance covariance matrix in the RWLLTM, MCMC outperforms MML estimation. The MCMC $\text{RMSE}_{\mathbf{D}}$ is signif-

Figure 1: $\text{RMSE}_{\mathbf{b}}$, $\text{RMSE}_{\boldsymbol{\beta}}$ and $\text{RMSE}_{\mathbf{D}}$ obtained with MCMC and MML estimation for the RWLLTM (top) and the LLTM (bottom)

icantly smaller than the corresponding estimation error for MML for 500 test takers and more. For the random coefficients a significant decrease in $\text{RMSE}_{\mathbf{b}}$ is even observed with MCMC for all sample sizes considered.

MCMC estimation for the RWLLTM is thus at least as good as MML estimation. Moreover, in case of many random feature effects and a large sample size MCMC estimation might even improve the estimation accuracy in comparison to MML. These results justify the use of MCMC as estimation technique in the remainder of the study.

4 Comparison study for the design criteria

In this section, minimum posterior weighted $\mathcal{D}$ -error ($\mathcal{D}_B$), minimum expected posterior weighted $\mathcal{D}$ -error ($E\mathcal{D}_B$), maximum expected Kullback-Leibler divergence between subsequent posteriors (KLP) and maximum mutual information (MUI) are compared as item selection rule to construct individual sequential designs for the RWLLTM. We consider test items incorporating 8 subtasks, either required or not to solve a specific item. Consequently, there are 256 ($= 2^8$) unique item feature combinations. Obviously, the item bank from which the items are selected includes (infinitely) many more test items. All feature effects are assumed random. Knowing that the RWLLTM also models a random intercept, the mean effects vector $\boldsymbol{\beta}$ and the individual effects vectors $\mathbf{b}_i$ are thus nine-dimensional. The true values for $\boldsymbol{\beta}$, $\mathbf{D}$ and all $\mathbf{b}_i$ were simulated and used, in turn, to simulate the responses for 1000 individuals with respect to the selected items in their tests. The coefficients in $\boldsymbol{\beta}$ were randomly drawn from $\mathcal{U}[-2, 2]$, the variance covariance matrix $\mathbf{D}$ was computed from a simulated Cholesky factor. The diagonal elements of $\mathbf{D}$ are between 0.223 and 3.135, the off-diagonals between -0.943 and 1.323. Each vector $\mathbf{b}_i$ was then sampled from $\mathcal{N}(\mathbf{0}, \mathbf{D})$.

The prior values $\boldsymbol{\beta}_0$ and $\mathbf{D}_0$ in the item selection criteria were set to respectively the zero vector and the identity matrix. Note that similar results were obtained with different prior information, assuring the robustness of the findings. The integrals in the design criteria were approximated with 1024 draws. For each individual a test design with 100 items was generated. As there are nine individual-specific coefficients in the model, a large amount of test data is

necessary to obtain sufficiently accurate estimates. Unfortunately, lengthy tests may induce learning or fatigue effects. In this study however, we assume that no item position effects are present, such that adaptive testing is warranted (Hohensinn et al., 2008; Kubinger, 2008). Further note that response simulation was repeated 50 times, the results reported in this section are averages over these repetitions.

From a smaller simulation study, including only 300 test takers and 30 items in their tests, it already became clear that ED_B is too complex, and therefore too slow, as item selection rule for the RWLLTM. Its computation time to select an additional item in a test highly exceeds the times of $\mathcal{D}_B$, KLP and MUI , making ED_B impractical for the RWLLTM. Moreover, ED_B did not yield more accurate estimates for the parameters in the model. Therefore, this design criterion was discarded from the larger study and results are only reported for $\mathcal{D}_B$, KLP and MUI .

Due to the role of the RWLLTM as cognitive diagnosis model, the main focus is on accurate estimation of the individual coefficients $\mathbf{b}_i$, representing the different dimensions of ability considered. To observe the change in estimation accuracy, the RWLLTM was not only estimated with all 100 items for each individual, but also using only the first 50, 60, 70, 80 and 90 test items. Figure 2 shows the mean $RMSE_{\mathbf{b}}$ values for each design criterion and each test length. Completely in line with expectations, a decrease in $RMSE_{\mathbf{b}}$, and therefore an increase in estimation accuracy for the individual coefficients, is observed for increasing test length. An intercriterion comparison on the other hand does not reveal any substantial differences between the item selection algorithms. No criterion stands out in efficiency or performs significantly better than the remaining criteria.

In addition to Figure 2, Figure 3 plots the estimates of b_i against the true values of the coefficients for all individuals obtained with either $\mathcal{D}_B$, KLP or MUI and test lengths 50 and 100. Note that the values in Figure 3 correspond to the eighth parameter. Plots for the remaining parameters are similar and therefore not displayed. The estimates obtained with the different item selection criteria are very alike, again illustrating the equivalence in design efficiency and estimation accuracy between the design algorithms.

Figure 2: Mean $\text{RMSE}_{\mathbf{b}}$ values obtained with $\mathcal{D}_B$, KLP and MUI and different test lengths

The plots in Figure 3 however also uncover some shrinkage of the estimated coefficients to the mean of zero. To estimate the random effects of an examinee not only the test data from that specific individual is used, but data and information from the entire population is incorporated. Estimates for the individual effects are thus, especially in case of few individual data, shrunk to the population mean. Nevertheless, even with only 50 items in the tests, the individual coefficients are being estimated quite accurately as the correlation between the estimates and the true values, averaged over the nine parameters, is respectively 0.9696, 0.9693 and 0.9690 for $\mathcal{D}_B$, KLP and MUI . Obviously, as already observed in Figure 2, estimation of the individual feature effects is improved as test length grows. Less shrinkage is observed in the right panel of Figure 3. Moreover, with 100 test items the average correlation between the estimates and the true values of the individual coefficients is increased to respectively 0.9834, 0.9834 and 0.9835.

Not only with respect to estimation of the individual coefficients, but also with respect to estimation of the population parameters $\boldsymbol{\beta}$ and $\mathbf{D}$ in the RWLLTM, the design criteria perform equally efficient. In Figure 4, the mean $\text{RMSE}_{\boldsymbol{\beta}}$ and mean $\text{RMSE}_{\mathbf{D}}$ values (with corresponding confidence intervals) are given for each design criterion and tests with 100 items. It is clear that the differences in average estimation error between the alternative design criteria are not significant. Similar conclusions could be drawn for the remaining test lengths.

Figure 3: Estimated parameters against the true values of b_{i8} for all individuals obtained with $\mathcal{D}_B$, KLP and MUI and test length 50 and 100

Figure 4: Mean $RMSE_{\beta}$ and $RMSE_D$ values obtained with $\mathcal{D}_B$, KLP and MUI and test length 100

In addition to the comparison of the design efficiency of the criteria, we have a closer look at the similarities and differences in the items they select by means of item exposure rates and an item overlap analysis. An exposure rate is computed for each unique item (more specifically, each unique item feature combination) in the item bank and is simply the percentage of test takers administered with that specific test item (Li and Schafer, 2005). They are given in Figure 5 for respectively $\mathcal{D}_B$, KLP and MUI and tests with 50, 70 and 100 items.

The exposure rates are very similar across the criteria. All three item selection rules apply all test items over the 1000 individualized designs (except for $\mathcal{D}_B$ and test length 50 for which one item remains unused). Due to differences in the (high dimensional) individual feature ef-

Figure 5: Exposure rates (%) for $\mathcal{D}_B$, KLP and MUI and test length 50, 70 and 100

	50 items			70 items			100 items	
	<i>KLP</i>	<i>MUI</i>		<i>KLP</i>	<i>MUI</i>		<i>KLP</i>	<i>MUI</i>
$\mathcal{D}_B$	34.51	34.10	$\mathcal{D}_B$	39.06	38.64	$\mathcal{D}_B$	45.39	44.95
<i>KLP</i>		34.91	<i>KLP</i>		39.19	<i>KLP</i>		45.23

Table 1: Item overlap (%) for test length 50, 70 and 100

fects, it makes sense that a large variety of test items is required to accurately estimate the many parameters in the RWLLTM. Further, the plots show no disturbing amounts of overused items. If a test item is selected and administered too often its content and solution become known among test takers. Correctly solving such items therefore no longer expresses individuals' true ability and the use of them could induce estimation bias. In this case, design algorithms should be adapted to account for item exposure control (Veldkamp and van der Linden, 2002).

But, even if two criteria select approximately the same test items over all designs, the items assigned to a specific individual can of course differ. To verify whether, and if so to what extent, the design criteria select the same items for the same test takers an item overlap analysis is conducted. The overlap rate between two design algorithms is the ratio of common items over the total amount of items in the test, averaged over individuals (Chen et al., 2000; Wang and Chang, 2011; Wang et al., 2011). The overlap rates between $\mathcal{D}_B$, *KLP* and *MUI* are given in Table 1 for test length 50, 70 and 100 and appear to be fairly high. For tests with 50 items, on average already more than one third of the test items are the same across the design criteria. The overlap is even approximately 45% for a test with 100 items.

From the simulations above, it can be concluded that $\mathcal{D}_B$, *KLP* and *MUI* are equivalently efficient as item selection rule to obtain test data and to estimate the RWLLTM as they yield equal estimation accuracy. *KLP* and *MUI* are however far less complex than the $\mathcal{D}_B$ criterion. The latter requires the computation of the determinant of the Fisher information matrix, which highly slows down the algorithm. For the simulation setup above, using 1024 draws for the approximation of the integrals in the design criteria, it takes in the beginning of the test on average slightly less than three seconds with the $\mathcal{D}_B$ criterion to select an additional item (Table 2). Although this seems acceptable, the difference with *KLP* and *MUI* is huge. Their computation time is approximately 0.242 seconds. Moreover, for all design criteria run times

Number of draws	$\mathcal{D}_B$	KLP	MUI
1024	2.742	0.242	0.243
2048	5.425	0.477	0.485
4096	10.713	0.930	0.953

Table 2: Average computation time (seconds) to select an additional test item with $\mathcal{D}_B$, KLP and MUI

increase as the test progresses because more and more data must be processed to select the next item. Note that the length of the test has an even larger effect on the computational cost of $\mathcal{D}_B$ as the Fisher information matrix includes all items in the test, slowing down its computation in case of longer tests. The distinction becomes even more dramatic when more draws are applied to compute the criteria. Table 2 shows that with 4096 draws it takes on average already more than ten seconds to select an item with $\mathcal{D}_B$, whereas for KLP and MUI still less than a second is required. As the selection of an additional item in a test takes on average more than 10 times longer with $\mathcal{D}_B$ than with KLP and MUI , the latter clearly outperform $\mathcal{D}_B$ in practicality.

5 Conclusion

The random weights linear logistic test model operates as a cognitive diagnosis model as it provides an instructive insight in the performance of examinees with respect to specific sub-tasks involved in test items. Models such as the two- and the three-parameter logistic model are still highly applied for test scoring, but they are only able to rank individuals according to their overall ability in a test’s subject. Recently however there has been a shift in interest to test models which are also able to give a more informative, multidimensional profile of one’s abilities.

To obtain valuable cognitive profiles, this research advocates an individualized design approach for the RWLLTM. Four item selection rules are discussed and compared: minimum posterior weighted ($\mathcal{D}_B$) and minimum expected posterior weighted ($E\mathcal{D}_B$) $\mathcal{D}$ -error, maximum expected Kullback-Leibler divergence between subsequent posteriors (KLP) and maximum mutual information (MUI). Due to an excessive complexity when applying it to the RWLLTM,

the ED_B criterion was however discarded from the simulation study.

The study could not reveal any significant differences in estimation accuracy between the different design criteria. Both the individual-specific coefficients and the population parameters in the RWLLTM are estimated equally well across the algorithms. $\mathcal{D}_B$, KLP and MUI thus all three appear useful to construct efficient individualized tests for the RWLLTM. The Kullback-Leibler criteria are however given preference due to their speed. The computation of KLP and MUI is far less complex, causing a huge decrease in computation time for an additional item in a test in comparison to $\mathcal{D}_B$. This study argues the relevance of the RWLLTM as cognitive diagnosis model and discusses different useful criteria to construct efficient individualized designs, as such it contributes to the expansion and improvement of recent developments in item response theory and optimal test design.

Appendix A: MCMC estimation for the RWLLTM

To obtain draws from the joint posterior distribution $[\boldsymbol{\beta}, \mathbf{D}, \mathbf{b} | \mathbf{Y}]$, the Gibbs sampler iteratively samples from the conditional posterior distributions: $[\boldsymbol{\beta} | \mathbf{D}, \mathbf{b}, \mathbf{Y}]$, $[\mathbf{D} | \boldsymbol{\beta}, \mathbf{b}, \mathbf{Y}]$ and $[\mathbf{b} | \boldsymbol{\beta}, \mathbf{D}, \mathbf{Y}]$ (Gamerman, 1997; Zeger and Karim, 1991). Fix starting values for $\boldsymbol{\beta}$, $\mathbf{D}$ and $\mathbf{b}_i \forall i$ to $\boldsymbol{\beta}^{(0)}$, $\mathbf{D}^{(0)}$ and $\mathbf{b}_i^{(0)} \forall i$. Given the t th draws $\boldsymbol{\beta}^{(t)}$, $\mathbf{D}^{(t)}$ and $\mathbf{b}_i^{(t)} \forall i$, the $(t + 1)$ th draws are obtained following subsequent steps.

$$(a) \quad [\boldsymbol{\beta} | \mathbf{D}^{(t)}, \mathbf{b}^{(t)}, \mathbf{Y}] = [\boldsymbol{\beta} | \mathbf{b}^{(t)}, \mathbf{Y}]$$

Assume prior $\phi(\boldsymbol{\beta} | \mathbf{a}, \mathbf{R})$ for $\boldsymbol{\beta}$ with ϕ the normal density. The conditional posterior $[\boldsymbol{\beta} | \mathbf{b}^{(t)}, \mathbf{Y}]$ of $\boldsymbol{\beta}$ is now proportional to

$$[\boldsymbol{\beta} | \mathbf{b}^{(t)}, \mathbf{Y}] \propto \prod_{i=1}^N \prod_{j=1}^J \frac{\exp[y_{ij}(\mathbf{x}'_j \boldsymbol{\beta} + \mathbf{z}'_j \mathbf{b}_i^{(t)})]}{1 + \exp(\mathbf{x}'_j \boldsymbol{\beta} + \mathbf{z}'_j \mathbf{b}_i^{(t)})} \exp \left[-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{a})' \mathbf{R}^{-1}(\boldsymbol{\beta} - \mathbf{a}) \right]. \quad (42)$$

A sample from the true distribution in (42) can be obtained through rejection sampling. Find $\boldsymbol{\beta}_M$, the posterior mode, and $\mathbf{H}_M$, the Hessian of the conditional posterior evaluated at the mode. Consider now a multivariate student t distribution with mean $\boldsymbol{\beta}_M$, variance

covariance matrix $-C_1 \times \mathbf{H}_M^{-1}$ and $\nu_{\boldsymbol{\beta}}$ degrees of freedom. Denote the multivariate t density by $g(\boldsymbol{\beta})$ and the target density in (42) by $f(\boldsymbol{\beta})$. In addition to C_1 , a constant C_2 is chosen such that $C_2 \cdot g(\boldsymbol{\beta})$ completely envelops the posterior, i.e. $C_2 \cdot g(\boldsymbol{\beta}) \geq f(\boldsymbol{\beta})$ over the range of $\boldsymbol{\beta}$. Following steps yield the $(t+1)$ th draw $\boldsymbol{\beta}^{(t+1)}$ from $f(\boldsymbol{\beta})$:

1. Randomly sample $\boldsymbol{\beta}^*$ from $g(\boldsymbol{\beta})$ and u from $\mathcal{U}(0, 1)$.
2. If $f(\boldsymbol{\beta}^*)/[C_2 \cdot g(\boldsymbol{\beta}^*)] \geq u$ then $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^*$, else return to the sampling step.

Parameters $\mathbf{a}$ and $\mathbf{R}$ in the prior for $\boldsymbol{\beta}$ were fixed to respectively $\mathbf{0}$ and $1000\mathbf{I}$, with $\mathbf{I}$ the identity matrix, $\nu_{\boldsymbol{\beta}}$ was set to 15.

(b) $[\mathbf{D}|\boldsymbol{\beta}^{(t)}, \mathbf{b}^{(t)}, \mathbf{Y}] = [\mathbf{D}|\mathbf{b}^{(t)}]$

Consider an inverted Wishart distribution with $\nu_{\mathbf{D}}$ degrees of freedom and scale matrix $\mathbf{S}$ (Gamerman, 1997; Lenk et al., 1996) as prior for $\mathbf{D}$. Generate $\mathbf{D}^{(t+1)}$ from an inverted Wishart distribution with $\nu_{\mathbf{D}} + N$ degrees of freedom and scale matrix $(\mathbf{S} + S^{(t)})/(\nu_{\mathbf{D}} + N)$, with

$$S^{(t)} = \sum_{i=1}^N \mathbf{b}_i^{(t)} (\mathbf{b}_i^{(t)})'. \quad (43)$$

A non-informative prior was assumed for $\mathbf{D}$, fixing $\nu_{\mathbf{D}}$ and $\mathbf{S}$ to respectively q and $\mathbf{I}$.

(c) $[\mathbf{b}_i|\boldsymbol{\beta}^{(t)}, \mathbf{D}^{(t)}, \mathbf{y}_i]$

The posterior for $\mathbf{b}_i$, conditional on the responses of individual i and previous draws of $\boldsymbol{\beta}$ and $\mathbf{D}$, is proportional to

$$[\mathbf{b}_i|\boldsymbol{\beta}^{(t)}, \mathbf{D}^{(t)}, \mathbf{y}_i] \propto L(\boldsymbol{\beta}^{(t)}, \mathbf{b}_i|\mathbf{y}_i, \mathbf{X}_i) \phi(\mathbf{b}_i|\mathbf{0}, \mathbf{D}^{(t)}) \quad (44)$$

$$\propto \prod_{j=1}^J \frac{\exp[y_{ij}(\mathbf{x}'_j \boldsymbol{\beta}^{(t)} + \mathbf{z}'_j \mathbf{b}_i)]}{1 + \exp(\mathbf{x}'_j \boldsymbol{\beta}^{(t)} + \mathbf{z}'_j \mathbf{b}_i)} \exp\left[-\frac{1}{2} \mathbf{b}_i' (\mathbf{D}^{(t)})^{-1} \mathbf{b}_i\right]. \quad (45)$$

There is no simple way to draw from this posterior, and so the Metropolis-Hastings algorithm (Train, 2003) is used.

1. Stack q independent standard normal values in a vector $\boldsymbol{\eta}$ and compute a trial value

$\tilde{\mathbf{b}}_i^{(t+1)} = \mathbf{b}_i^{(t)} + \rho L \boldsymbol{\eta}$, with L the Cholesky factor of $\mathbf{D}^{(t)}$. Note that ρ is updated here

throughout the algorithm in order to maintain an acceptance rate of approximately 30% over individuals.

2. Compute the ratio

$$F = \frac{L(\boldsymbol{\beta}^{(t)}, \tilde{\mathbf{b}}_i^{(t+1)} | \mathbf{y}_i, \mathbf{X}_i) \phi(\tilde{\mathbf{b}}_i^{(t+1)} | \mathbf{0}, \mathbf{D}^{(t)})}{L(\boldsymbol{\beta}^{(t)}, \mathbf{b}_i^{(t)} | \mathbf{y}_i, \mathbf{X}_i) \phi(\mathbf{b}_i^{(t)} | \mathbf{0}, \mathbf{D}^{(t)})}. \quad (46)$$

3. Draw a value u from $\mathcal{U}(0, 1)$. If $F \geq u$ then accept the trial $\mathbf{b}_i^{(t+1)} = \tilde{\mathbf{b}}_i^{(t+1)}$. In case $F < u$, $\mathbf{b}_i^{(t+1)} = \mathbf{b}_i^{(t)}$.

For a large amount of iterations the draws converge to the joint distribution of $\boldsymbol{\beta}, \mathbf{D}$ and $\mathbf{b}_i \forall i$. Estimates of the model parameters are obtained by taking means of the draws.

In the simulation study comparing MCMC with MML estimation, 15000 iterations were performed of which 10000 used for burn-in and the remaining 5000, without tinning, for parameter estimation. In the study comparing the design criteria the number of iterations is 100000, of which 70000 burn-in draws. The remaining draws were tinned with every 10th value. In all simulations only one set of starting values was applied, more specifically the zero vector for $\boldsymbol{\beta}$ and all $\mathbf{b}_i$ and the identity matrix for $\mathbf{D}$. Due to the large amounts of iterations and the long burn-ins, different starting values did not alter the final estimates.

Appendix B: Importance sampling

To apply the design criteria in practice, we approximate the integrals over updated posteriors through importance sampling. Assume we want to compute

$$C = \int C(\mathbf{b}) f(\mathbf{b} | \mathbf{y}_{k-1}) d\mathbf{b}, \quad (47)$$

which weighs a criterion C depended on $\mathbf{b}$ over the posterior $f(\mathbf{b} | \mathbf{y}_{k-1})$ obtained after observing the responses to $k - 1$ items. Importance sampling proceeds as follows:

1. Find the posterior mode $\mathbf{b}_M$ and the Hessian $\mathbf{H}_M$ of the posterior distribution $f(\mathbf{b} | \mathbf{y}_{k-1})$ evaluated at the mode and take as importance density the multivariate student t dis-

tribution with mean $\mathbf{b}_M$, variance covariance matrix $-\mathbf{H}_M^{-1}$ and ν , here 8, degrees of freedom.

2. Sample a large amount R of draws $\mathbf{b}^r$ from the importance density and construct for each draw the importance weight w^r . Consider $f^*(\mathbf{b}|\mathbf{y}_{k-1})$ and $\psi(\mathbf{b})$ the kernel of respectively the posterior distribution and the importance density. Weight w^r equals

$$w^r = \frac{f^*(\mathbf{b}^r|\mathbf{y}_{k-1})/\psi(\mathbf{b}^r)}{\sum_{s=1}^R f^*(\mathbf{b}^s|\mathbf{y}_{k-1})/\psi(\mathbf{b}^s)}. \quad (48)$$

Note that this research considers extensible shifted lattice points with Baker's transformation (Yu et al., 2010), instead of random draws, to approximate the integrals.

3. The criterion in (47) is then approximated by

$$\tilde{C} = \sum_{r=1}^R C(\mathbf{b}^r) \times w^r. \quad (49)$$

References

- Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum experimental designs, with SAS. Clarendon Press, Oxford U.K.
- Berger, M., 1992. Sequential sampling designs for the two-parameter item response theory model. *Psychometrika*. 57, 521-538.
- Bouwmeester, S., van Rijen, E.H.M., Sijtsma, K., 2011. Understanding phoneme segmentation performance by analyzing abilities and word properties. *European Journal of Psychological Assessment*. 27, 95-102.
- Chang, H.H., Ying, Z., 1996. A global information approach to computerized adaptive testing. *Applied Psychological Measurement*. 20, 213-229.

Chen, S.Y., Ankenmann, R.D., Chang, H.H., 2000. A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*. 24, 241-255.

Chen, Y.H., MacDonald, G., Leu, Y.C., 2011. Validating cognitive sources of mathematics item difficulty: application of the LLTM to fraction conceptual items. *The International Journal of Educational and Psychological Assessment*. 7, 74-93.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., Partchev, I., 2011. The estimation of item response models with the `lmer` function from the `lme4` package in R. *Journal of Statistical Software*. 39, 1-28.

de la Torre, J., 2009. A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*. 33, 163-183.

de la Torre, J., Douglas, J.A., 2008. Model evaluation and multiple strategies in cognitive diagnosis: an analysis of fraction subtraction data. *Psychometrika*. 73, 595-624.

Embretson, S.E., Daniel, R.C., 2008. Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly*. 50, 328-344.

Fischer, G.H., 1973. The linear logistic test model as an instrument in educational research. *Acta Psychologica*. 37, 359-374.

Freund, P.A., Hofer, S., Holling, H., 2008. Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological*

Measurement. 32, 195-210.

Gamerman, D., 1997. Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*. 7, 57-68.

Hahne, J., 2008. Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*. 50, 379-390.

Henson, R., Douglas, J., 2005. Test construction for cognitive diagnosis. *Applied Psychological Measurement*. 29, 262-277.

Hohensinn, C., Kubinger, K.D., 2009. On varying item difficulty by changing the response format for a mathematical competence test. *Austrian Journal of Statistics*. 38, 231-239.

Hohensinn, C., Kubinger, K.D., 2011. Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement*. 71, 732-746.

Hohensinn, C., Kubinger, K.D., Reif, M., Holocher-Ertl, S., Khorramdel, L., Frebort, M., 2008. Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*. 50, 391-402.

Holling, H., Bertling, J.P., Zeuch, N., 2009. Automatic item generation of probability word problems. *Studies in Educational Evaluation*. 35, 71-76.

Holling, H., Blank, H., Kuchenbäcker, K., Kuhn, J.T., 2008. Rule-based item design of statistical word problems: a review and first implementation. *Psychology Science Quarterly*.

terly. 50, 363-378.

Hornke, L.F., Habon, M.W., 1986. Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*. 10, 369-380.

Kubinger, K.D., 2008. On the revival of the Rasch model-based LLTM: from constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly*. 50, 311-327.

Kubinger, K.D., 2009. Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*. 69, 232-244.

Kubinger, K.D., Hohensinn, C., Holocher-Ertl, S., Heuberger, N., 2011. Applying the LLTM for the determination of children's cognitive age-acceleration function. *Psychological Test and Assessment Modeling*. 53, 183-191.

Lenk, P.J., DeSarbo, W.S., Green, P.E., Young, M.R., 1996. Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*. 15, 173-191.

Li, Y.H., Schafer, W.D., 2005. Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement*. 29, 3-25.

McGlohen, M., Chang, H.H., 2008. Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*. 40, 808-821.

Medina-Diaz, M., 1993. Analysis of cognitive structure using the linear logistic test

model and quadratic assignment. *Applied Psychological Measurement*. 17, 117-130.

Mulder, J., van der Linden, W.J., 2009. Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*. 74, 273-296.

Mulder, J., van der Linden, W.J., 2010. Multidimensional adaptive testing with Kullback-Leibler information item selection. In: van der Linden, W.J., Glas, C.A.W. (eds.) *Elements of adaptive testing*. pp. 77-101. Springer, New York.

Poinstingl, H., 2009. The linear logistic test model (LLTM) as the methodological foundation of item generating rules for a new verbal reasoning test. *Psychology Science Quarterly*. 51, 123-134.

Rijmen, F., De Boeck, P., 2002. The random weights linear logistic test model. *Applied Psychological Measurement*. 26, 271-285.

Segall, D.O., 2010. Principles of multidimensional adaptive testing. In: van der Linden, W.J., Glas, C.A.W. (eds.) *Elements of adaptive testing*. pp. 57-75. Springer, New York.

Sonnleitner, P., 2008. Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*. 50, 345-362.

Train, K., 2003. *Discrete choice methods with simulation*. Cambridge University Press.

van der Linden, W.J., 1998. Bayesian item selection criteria for adaptive testing. *Psychometrika*. 63, 201-216.

van der Linden, W.J., Pashley, P.J., 2010. Item selection and ability estimation in

adaptive testing. In: van der Linden, W.J., Glas, C.A.W. (eds.) Elements of adaptive testing. pp. 3-30. Springer, New York.

Veldkamp, B.P., van der Linden, W.J., 2002. Multidimensional adaptive testing with constraints on test content. *Psychometrika*. 67, 575-588.

Wang, C., Chang, H.H., 2011. Item selection in multidimensional computerized adaptive testing - gaining information from different angles. *Psychometrika*. 76, 363-384.

Wang, C., Chang, H.H., Boughton, K.A., 2011. Kullback-Leibler information and its applications in multidimensional adaptive testing. *Psychometrika*. 76, 13-39.

Weissman, A., 2007. Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*. 67, 41-58.

Whitely, S.E., Schneider, L.M., 1981. Information structure for geometric analogies: a test theory approach. *Applied Psychological Measurement*. 5, 383-397.

Yao, L., Boughton, K.A., 2007. A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*. 31, 83-105.

Yu, J., Goos, P., Vandebroek, M., 2010. Comparing different sampling schemes for approximating the integrals involved in the efficient design of stated choice experiments. *Transportation Research Part B*. 44, 1268-1289.

Yu, J., Goos, P., Vandebroek, M., 2011. Individually adapted sequential Bayesian conjoint-choice designs in the presence of consumer heterogeneity. *International Journal of*

Research in Marketing. 28, 378-388.

Zeger, S.L., Karim, M.R., 1991. Generalized linear models with random effects: a Gibbs sampling approach. Journal of the American Statistical Association. 86, 79-86.

Zeuch, N., Holling, H., Kuhn, J.T., 2011. Analysis of the latin square task with linear logistic test models. Learning and Individual Differences. 21, 629-632.

FACULTY OF ECONOMICS AND BUSINESS

Naamsestraat 69 bus 3500

3000 LEUVEN, BELGIË

tel. + 32 16 32 66 12

fax + 32 16 32 67 91

info@econ.kuleuven.be

www.econ.kuleuven.be