An efficient standard error estimator of the DINA model parameters when analysing clustered data

Jung Yeon Park*

Faculty of Psychology and Educational Sciences,
Katholieke Universiteit Leuven,
Etienne Sabbelaan 51 - box 7800,
8500 Kortrijk, Belgium
Email: ellie.park@kuleuven.be
*Corresponding author

Young-Sun Lee and Matthew S. Johnson

Department of Human Development at Teachers College
Columbia University,
525 West 120th St.,
New York, NY 10027, USA
Email: yslee@exchange.tc.columbia.edu
Email: Johnson@exchange.tc.columbia.edu

Abstract: Cognitive diagnostic modelling is often used to analyse educational and psychological data, which are typically collected through cluster sampling with unequal selection probabilities. Jackknife is a resampling technique used to account for the sampling design. It typically gives unbiased estimates of the standard errors of the model parameters, but implementation can be vastly time-consuming. This study proposes an accurate and computationally fast approach for the standard errors of the parameters in the DINA model, one that incorporates the Huber–White sandwich estimator approach. Our simulation study suggests that the proposed sandwich estimator performs well when analysing clustered data structures specifically with moderate to large numbers of clusters. We also demonstrate its applicability to TIMSS 2011 mathematics.

Keywords: cluster sampling; DINA; generalised estimating equation; jackknife resampling; sandwich estimator.

Reference to this paper should be made as follows: Park, J.Y., Lee, Y-S. and Johnson, M.S. (2017) 'An efficient standard error estimator of the DINA model parameters when analysing clustered data', *Int. J. Quantitative Research in Education*, Vol. 4, Nos. 1/2, pp.159–190.

Biographical notes: Jung Yeon Park is a Postdoctoral Researcher in imec - ITEC - KU Leuven at the University of Leuven, Belgium. Her research interests include cognitive diagnostic modelling, item response theory, and computerised adaptive learning.

Young-Sun Lee is an Associate Professor of Psychology and Education in the Department of Human Development at Teachers College, Columbia University. Her primary research interests include applications of item response theory and cognitive diagnostic modelling.

Matthew S. Johnson is an Associate Professor of Statistics and Education in the Department of Human Development at Teachers College, Columbia University. His research interests include educational statistics, item response theory, Bayesian statistics and educational testing.

1 Introduction

Cognitive diagnostic models (CDMs) are a type of latent class models (LCMs; Dayton and Macready, 1988; Haertel, 1989) used when the successful performance of a task is parameterised by an examinee's possession or lack of fine-grained discrete skills. A wide range of models fall within the framework of CDMs. Examples include deterministic-input, noisy-and-gate (DINA; Junker and Sijtsma, 2001) as a conjunctive model, deterministic-input, noisy-or-gate (Templin and Henson, 2006) as a compensatory model, plus more generalised types of the diagnostic models (e.g. de la Torre, 2011; Henson et al., 2009; von Davier, 2005). Among the choices, the DINA model has elicited the most interest from researchers and practitioners owing to its analytical tractability and conceptual interpretability (de la Torre, 2011).

Cognitive diagnostic models have been utilised in real-world data analyses including large-scale educational assessment data (Chen and Chen, 2015; Chen and de la Torre, 2014; Johnson et al., 2013; Lee et al., 2011; Xu and von Davier, 2008). Typically, such empirical data are collected from large-scale student populations, with clusters naturally occurring due to the sampling procedures - e.g. students are nested within geographic areas or schools. In this case students within the same geographic area tend to share a similar learning propensity, and the students' outcomes may be correlated. Ignoring the correlations among outcomes that arise due to the nature of the clustered structure can produce invalid variance estimates and thereby misleading statistical inference and testing of various psychometric properties; e.g. Differential Item Functioning (DIF) detection procedure (Hou et al., 2014) and model comparisons (de la Torre, 2011; de la Torre and Lee, 2013). To handle the issue, pseudo-replication technique (i.e. jackknife (JK) or bootstrap) is commonly used (Wolter, 2007). More specifically, the jackknife procedure (Quenouille, 1949) is widely applied so as to address the cluster sampling designs for a wide range of the LCMs. Patterson et al. (2002), e.g. assessed the accuracy of the jackknife technique to account for stratified cluster sampling in estimating the standard errors (SEs) of the latent class (LC) probabilities of the traditional LCM. The authors concluded that the technique showed sufficiently accurate performance across many empirical analyses, although it tended to "slightly overestimate the actual standard errors." They also found that ignoring the sampling weight had a significant impact on estimating the standard errors of the LC probabilities.

The jackknife is often adapted to the CDMs for analysing large-scale data from the cluster sampling. Hsieh et al. (2010), e.g. when addressing the sampling design of the NAEP, used the method to calculate standard error of the latent ability parameter within the general diagnostic model (GDM; von Davier, 2005). Johnson et al. (2013) used the technique to estimate a covariance matrix for the latent skill pattern probabilities of the multiple-group DINA model in analysing Trends in International Mathematics and Science Study (TIMSS) for the purpose of comparing the skill distributions of multiple countries. Despite the jackknife's applicability to the CDMs, however, its accuracy has

not been fully evaluated across the realistic simulation conditions of the sampling designs. More importantly, the technique cannot obviate its fundamental limitation: computational inefficiency due to the repeated resampling (de Leeuw et al., 2008; Johnson et al., 1990). Implementation occasionally suffers from a long computational time, as the numbers of clusters and cluster sizes increase.

An alternative approach to estimating standard error in the presence of the cluster sampling is that of using the Taylor series expansion of the appropriate estimating equations (Wolter, 2007). The asymptotic variance of the linear terms in the Taylor series yields a consistent estimator of the variance of the statistic, namely the Huber–White sandwich variance estimator (Huber, 1967, 1981; White, 1982). The generalised estimating equation (GEE; Liang and Zeger, 1986) is a set of estimating equations commonly used for clustered data. The GEE-based sandwich estimator uses the covariance structure of the repeated outcomes, and this approach leads to an efficient estimator even when the working covariance structure is misspecified. Often it is used in longitudinal studies by modelling the correlations between outcomes across multiple time points, but it can also model various other types of repeated outcomes (Fitzmaurice et al., 2004). For example, Ip and Chen (2012) constructed a modified GEE-based sandwich estimator by modelling correlations between item responses. They noted that assuming local independence for the unidimensional Item Response Theory (IRT) is often unrealistic or too stringent (Yen, 1993; Ip, 2000, 2010). In this case using the naïve approach can yield a biased standard error for the latent ability parameter, particularly when there is a large number of test items. The authors have demonstrated that the proposed sandwich estimator can efficiently adjust for the impact of conditional correlations among the item responses.

The goal of the present study is to develop efficient standard error (or variance) estimators of the DINA model in the presence of a clustered data structure. Using GEE we derive the sandwich estimators for the variances of the guess, slip and skill probability parameters of the DINA model. The proposed approach aims to:

a match or even improve upon the accuracy of the jackknife estimator

b be speedier than the resampling technique.

The rest of the paper is organised as follows. The next two sections (Sections 2 and 3) provide a brief overview of the DINA model framework and the basic idea of using the sandwich formulation for a broad class of estimators. We then propose two modified versions of the sandwich estimators originally proposed by Liang and Zeger (1986) and Pan (2001) to accommodate cluster effects in estimating DINA model parameters. A simulation study (Section 4) compares the performance of the jackknife method to that of the proposed estimators when there are various numbers of clusters and cluster sizes. In Section 5, these methods are used to estimate the standard errors of the parameters in the TIMSS 2011 mathematics data. In Section 6, we briefly demonstrate applicability to GDM. In Section 7, we discuss our findings.

2 The DINA model

Let us suppose that the data contain the responses of N examinees to J items in a test. A total of J item responses ($j = 1,\ldots,J$) is measured for each examinee i ($i = 1,\ldots,N$); that is, $Y_{ij} = 1$ if the examinee i answers item j correctly, $Y_{ij} = 0$ otherwise. We further let

$\boldsymbol{Y}_i = (y_{i1},\ldots,y_{iJ})^T$ denote the observed response pattern of the examinee *i*. Next, suppose there is a latent skill parameter, α_k, signifying mastery or non-mastery of each knowledge skill for each examinee. More specifically, $\alpha_{ik} = 1$ indicates that the examinee *i* has mastered the skill *k* ($k = 1,\ldots,K$). Similarly, $\alpha_{ik} = 0$ indicates that the examinee *i* has not mastered the skill *k*. Because the individual skill parameters, $\alpha_1 \ldots \alpha_K$, are all binary indicators, there exist 2^K mutually exclusive skill patterns, $\alpha_0 \ldots \alpha_l \ldots \alpha_L$, where $l = 0,\ldots,L; L = (2^K - 1)$. Here, α_0 denotes the pattern when all skills are absent, i.e. $\alpha_0 = (0,\ldots,0)^T$; α_L denotes the pattern when all skills are present, i.e. $\alpha_L = (1,\ldots,1)^T$. Thus, the latent skill space can be described by the probabilities of each of the skill patterns, $\boldsymbol{\pi}_A = (\pi_0,\ldots,\pi_l,\ldots,\pi_L)^T$.

The DINA model (Junker and Sijtsma, 2001) is characterised by a conjunctive rule for the latent response (=η_{ij}), to determine the examinee *i*'s task performance given one's skill patterns. In other words, $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, which means that one must master all of the required skills to correctly answer item *j*. Here, the q_{jk} is a binary indicator used to specify whether skill k is needed to correctly answer the item *j*, i.e. $q_{jk} \in \{0,1\}$, while the matrix of all q_{jk}'s is called the 'Q-matrix' (Tatsuoka, 1985). The latent response η_{ij} is linked to the observed responses within a probabilistic relationship having two 'noisy' parameters - i.e. slip (=s_j) and guess (=g_j) parameters. The guess rate is the probability that the examinee *i* has answered item *j* correctly, even though one does not possess all of the required skills - i.e. $g_j = P(Y_{ij} = 1 \mid \eta_{ij} = 0)$. The slip rate is the probability that the examinee *i* fails to respond to item *j* correctly even though one possesses all of the required skills - i.e. $s_j = P(Y_{ij} = 0 \mid \eta_{ij} = 1)$. Putting together, the item response function for the DINA model is written as

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_l) = (1 - s_j)^{\eta_{ij}} \, g_j^{1-\eta_{ij}}. \tag{1}$$

In estimating the DINA model parameters, the marginal maximum likelihood estimation method can be used. Assuming that (s_j, g_j) and $\boldsymbol{\alpha}_l$ are independent, the marginal probability for answering item *j* correctly - i.e. $\mu_{ij} = P(\boldsymbol{Y}_{ij} = 1)$ - is as follows:

$$P(\boldsymbol{Y}_{ij} = 1) = \sum_{l=0}^{L} \pi_l P(\boldsymbol{Y}_{ij} = 1 \mid \boldsymbol{\alpha}_l), \tag{2}$$

where π_l denotes a probability of the skill pattern *l*. Assuming that all Y_{ij}'s are conditionally independent for all *j*'s given $\boldsymbol{\alpha}_l$ and moderately correlated within each cluster, the following sample-weighted pseudo log-likelihood can be used, with the sampling weights then being incorporated via the specific sample selection process:

$$\log P(\boldsymbol{Y}) = \sum_i w_i \log \sum_{l=0}^{L} \pi_l P(\boldsymbol{Y}_i \mid \boldsymbol{\alpha}_l), \tag{3}$$

where $P(\boldsymbol{Y}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J}\left[s_j^{1-y_{ij}}(1-s_j)^{y_{ij}}\right]^{\eta_{ij}}\left[g_j^{y_{ij}}(1-g_j)^{1-y_{ij}}\right]^{1-\eta_{ij}}$ and w_i indicates the sampling weight of the examinee *i*. As for incorporating the sampling weight for latent variable models, using *examinee*-level weight is conventional across many literatures (Patterson et al., 2002; Rabe-Hesketh and Skrondal, 2006; Wedel et al., 1998).

3 The sandwich standard error

Let us recall that $\boldsymbol{Y}_i = (y_{i1},\ldots,y_{iJ})^T$ denotes a vector of *J* repeated responses ($j = 1,\ldots,J$) by the examinee *i* ($i = 1,\ldots,N$). Suppose a vector of mean responses, $\boldsymbol{\mu}_i = (\mu_{i1},\ldots,\mu_{iJ})^T$, where μ_{ij} is a function of a set of parameters, say $\boldsymbol{\beta}$. From the quasi-likelihood estimation theory (Wedderburn, 1974), the following generalised estimation equation yields the consistent estimator $\hat{\boldsymbol{\beta}}$ for the $\boldsymbol{\beta}$:

$$\sum_i \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = 0, \tag{4}$$

where $\boldsymbol{D}_i$ denotes the derivative matrix of $\boldsymbol{\mu}_i$ with respect to the $\boldsymbol{\beta}$; $\boldsymbol{S}_i = (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$; $\boldsymbol{V}_i$ denotes a working covariance matrix such that $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}\boldsymbol{A}_i^{1/2}$, where $\boldsymbol{A}_i = \mathrm{diag}(v(\mu_{i1}),\ldots,v(\mu_{iJ}))$ and *v* is a known variance function of μ_i. In the case of binary responses, $v(\mu_{ij}) = \mu_{ij}(1-\mu_{ij})$. Finally, the $\boldsymbol{R}$ in the $\boldsymbol{V}_i$ denotes a working correlation that we separately model so as to represent a specific type of within-subject correlation structure.

In the above setup, and under mild regularity conditions, $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically multivariate normal with zero mean and the variance estimator $\Sigma(\hat{\boldsymbol{\beta}})$ for $\hat{\boldsymbol{\beta}}$ as follows (see the Proof of Theorem 2 from Liang and Zeger, 1986):

$$\Sigma(\hat{\boldsymbol{\beta}}) = \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}, \tag{5}$$

where

$$\boldsymbol{B} = \sum_i \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i, \text{ and} \tag{6}$$

$$\boldsymbol{M} = \sum_i \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \mathrm{Cov}(\boldsymbol{Y}_i) \boldsymbol{V}_i^{-1} \boldsymbol{D}_i. \tag{7}$$

The $\mathrm{Cov}(\boldsymbol{Y}_i)$ in Eq. (7) indicates the true variance-covariance matrix of $\boldsymbol{Y}_i$. The sandwich variance estimator leads to a consistent variance of the $\Sigma(\hat{\boldsymbol{\beta}})$ even when the correlation structure of the repeated measurements is misspecified. When data are collected via pure simple random sampling and when the working covariance, $\boldsymbol{V}_i$, is correctly specified - i.e. $\mathrm{Cov}(\boldsymbol{Y}_i) = \boldsymbol{V}_i$ - then $\Sigma(\hat{\boldsymbol{\beta}})$ is reduced to $\boldsymbol{B}^{-1}$.

3.1 Variance covariance structure of Y_i

In the present study, we adopt the two variance estimators constructed respectively by Liang and Zeger (1986) and Pan (2001) so as to derive the closed-form solutions and thereby obtain two variance formulas for the DINA model parameters. First, according to Liang and Zeger (1986), the variance-covariance matrix of $\boldsymbol{Y}_i$ can be approximated as follows:

$$\mathrm{Cov}\left(\boldsymbol{Y}_i\right)_{\mathbf{LZ}} = \boldsymbol{S}_i\boldsymbol{S}_i^T = \left(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i\right)\left(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i\right)^T, \tag{8}$$

where it is estimated for each individual examinee i. $\hat{\boldsymbol{\mu}}_i$ for the vector of the mean responses is obtained by using the estimator $\hat{\boldsymbol{\beta}}$.

In order to improve the performance of the variance estimator, Pan (2001) proposed an alternative method of estimating the true variance–covariance matrix of $\boldsymbol{Y}_i$. Given that there is a common correlation structure across n examinees, $\mathrm{Cov}(\boldsymbol{Y}_i)$ is calculated by *pooling* all examinees i^*'s ($i^* = 1,\ldots,n$) in a sample as follows:

$$\mathrm{Cov}\left(\boldsymbol{Y}_i\right) = \boldsymbol{A}_i^{1/2}\left(\frac{1}{n}\sum_{i^*=1}^{n}\boldsymbol{A}_{i^*}^{-1/2}\boldsymbol{S}_{i^*}\boldsymbol{S}_{i^*}^T\boldsymbol{A}_{i^*}^{-1/2}\right)\boldsymbol{A}_i^{1/2}. \tag{9}$$

In this study, we modify the variance–covariance matrix of $\boldsymbol{Y}_i$ in Eq. (9), for the purpose of incorporating the sampling design effect.

Let $T\left(t = 1,\ldots,T\right)$ be the number of clusters in the sample representing for a population, n_t be the number of individuals within each cluster t and w_i be the examinee-specific sampling weight for the examinee i. The GEE-based sandwich variance estimator, based on the top-level clusters, is usually able to account for the multilevel correlation structures if the multiple clusters are perfectly nested (Betensky et al., 2000). Note that we modified the pooled variance–covariance matrix formula in Eq. (9), by generating 'cluster-specific' covariance matrices from the pooled examinees in the cluster. For any examinee i sampled from the cluster t, let i_t^* denote examinees within the cluster t, $i_t^* = 1,\ldots,n_t$. Then,

$$\begin{aligned}\mathrm{Cov}\left(\boldsymbol{Y}_i\right)_{\mathbf{Pan}} &= \boldsymbol{A}_i^{1/2}\left(\frac{1}{n_t}\sum_{i_t^*=1}^{n_t}\boldsymbol{A}_{i_t^*}^{-1/2}\boldsymbol{S}_{i_t^*}\boldsymbol{S}_{i_t^*}^T\boldsymbol{A}_{i_t^*}^{-1/2}\right)\boldsymbol{A}_i^{1/2} \\ &= \boldsymbol{A}_i^{1/2}\left(\frac{1}{n_t}\sum_{i_t^*=1}^{n_t}\boldsymbol{A}_{i_t^*}^{-1/2}\left(\boldsymbol{Y}_{i_t^*} - \hat{\boldsymbol{\mu}}_{i_t^*}\right)\left(\boldsymbol{Y}_{i_t^*} - \hat{\boldsymbol{\mu}}_{i_t^*}\right)^T\boldsymbol{A}_{i_t^*}^{-1/2}\right)\boldsymbol{A}_i^{1/2}\end{aligned}, \tag{10}$$

where $\boldsymbol{S}_{i_t^*} = \left(\boldsymbol{Y}_{i_t^*} - \hat{\boldsymbol{\mu}}_{i_t^*}\right)$ and $\boldsymbol{A}_{i_t^*} = \hat{\boldsymbol{\mu}}_{i_t^*}\left(1 - \hat{\boldsymbol{\mu}}_{i_t^*}\right)$.

3.2 Working correlation matrix

In modelling the working correlation matrix, $\boldsymbol{V}_i$ in Eqs. (6) and (7), various types of correlation matrices can be used depending upon the researcher's need (Fitzmaurice

et al., 2004). In the present case, we are interested in a variance estimator capable of accounting for possibly different correlation structures between clusters, hence our decision to use separate working correlation matrices for each cluster. The proposed working correlation matrix assumes that there is a unique correlation structure shared by all of the examinees within a cluster. We also assume that the correlation matrices are unstructured. Therefore, Pearson correlation for the examinees in a cluster is used to estimate the corresponding working correlation matrices for the cluster; $\boldsymbol{V}_t$, where $t = 1,\ldots,T$. The next two sections demonstrate how the partial derivative matrices for the item parameters and skill probability are obtained.

3.3 Derivative matrices for guess and slip parameters

Let us suppose a matrix of partial derivatives $\boldsymbol{D}_i$ of $\boldsymbol{\mu}_i = \left(\mu_{i1},\ldots,\mu_{iJ}\right)^T$ in Eqs. (4)–(7), with respect to the two components of the item parameters - i.e. $\boldsymbol{g} = \left(g_1,\ldots,g_J\right)^T$ for guess, and $\boldsymbol{s} = \left(s_1,\ldots,s_J\right)^T$ for slip as follows:

$$\boldsymbol{D}_i = \left[\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{g}} \vdots \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{s}}\right]^T, \tag{11}$$

where the subset of the matrix with respect to the guess and slip parameters comprises ($J \times J$) elements as follows:

$$\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{g}} = \begin{bmatrix} \frac{\partial \mu_{i1}}{\partial g_1} & \cdots & \frac{\partial \mu_{iJ}}{\partial g_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_{i1}}{\partial g_J} & \cdots & \frac{\partial \mu_{iJ}}{\partial g_J} \end{bmatrix} \quad \text{and} \quad \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{s}} = \begin{bmatrix} \frac{\partial \mu_{i1}}{\partial s_1} & \cdots & \frac{\partial \mu_{iJ}}{\partial s_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_{i1}}{\partial s_J} & \cdots & \frac{\partial \mu_{iJ}}{\partial s_J} \end{bmatrix}. \tag{12}$$

We recall from Section 2 that $\eta_j(\boldsymbol{\alpha}) = 1$ denotes the skill pattern $\boldsymbol{\alpha}$ possesses all the skills required to answer item *j* correctly, while $\eta_j(\boldsymbol{\alpha}) = 0$ denotes the skill pattern $\boldsymbol{\alpha}$ lacks at least one of the skills required to answer item *j* correctly. Then $\mu_{ij} = P\left(Y_{ij} = 1\right)$ denotes the probability of giving the correct response for the examinee *i* to the item *j*, which can be decomposed by combining the item parameters (i.e. g_j and s_j) and the skill pattern probability (i.e. π_l, $l = 0,\ldots,2^{K-1}$) as follows:

$$\begin{aligned} \mu_{ij} &= P(Y_{ij} = 1) = \sum_{l=0}^{L} \pi_l P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_l) \\ &= g_j \sum_{\boldsymbol{\alpha}_l \in A^{(-)}} \pi_l + (1 - s_j) \sum_{\boldsymbol{\alpha}_l \in A^{(+)}} \pi_l \end{aligned}, \tag{13}$$

where $A^{(-)} = \{\boldsymbol{\alpha}_l \mid \eta_j(\boldsymbol{\alpha}_l) = 0\}$ and $A^{(+)} = \{\boldsymbol{\alpha}_l \mid \eta_j(\boldsymbol{\alpha}_l) = 1\}$. As discussed in Section 2, conditional probability $P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_l)$ in Eq. (13) can be divided into two cases:

1 guessing item *j* correctly, even when some of the required skills are absent

2 not slipping to answer item *j* correctly, because all the skills for the item are present.

In Eq. (13), $\sum_{\boldsymbol{\alpha}_l \in A^{(-)}} \pi_l$ denotes a summation of all probabilities when the examinee *i* lacks at least one of the skills required to solve item *j* correctly. Similarly, $\sum_{\boldsymbol{\alpha}_l \in A^{(+)}} \pi_l$ denotes a summation of all probabilities, with the skill patterns possessing all skills required to solve item *j* correctly. Given Eq. (13) for μ_{ij}, each element of the partial derivative matrix, with respect to the guess parameter for item *j*, can be divided into two scenarios:

$$\frac{\partial \mu_{ij}}{\partial g_{j'}} = \begin{cases} \sum_{\alpha_l \in A^{(-)}} \pi_l & j = j' \\ 0 & j \neq j' \end{cases}. \tag{14}$$

Similarly, each element of the matrix with respect to the slip parameter for item *j* is as follows:

$$\frac{\partial \mu_{ij}}{\partial s_{j'}} = \begin{cases} -\sum_{\alpha_l \in A^{(+)}} \pi_l & j = j' \\ 0 & j \neq j' \end{cases}. \tag{15}$$

The above partial derivative matrix $\boldsymbol{D}_i$ of the guess and slip parameters is, thereby, incorporated to the bread and meat portions in Eqs. (6) and (7). With all the results from Sections 3.1–3.3 incorporated to Eqs. (6) and (7), the sandwich covariance matrix $\Sigma(\hat{\boldsymbol{g}})$ and $\Sigma(\hat{\boldsymbol{s}})$ in Eq. (5) for the item parameters can be obtained. Finally, the square root of the diagonal elements of $\Sigma(\hat{\boldsymbol{g}})$ and $\Sigma(\hat{\boldsymbol{s}})$ in Eq. (5) results in the sandwich SEs of the guess and slip parameters corresponding to each item.

3.4 Derivative matrix for skill probability

Let us denote the vector of skill pattern probabilities as $\boldsymbol{\pi}_{\boldsymbol{A}} = (\pi_0, \ldots, \pi_l, \ldots, \pi_L)^T$, where π_l is the probability of the *l*th skill pattern ($=\boldsymbol{\alpha}_l$), $l = 0, \ldots, L$; $L = (2^K - 1)$. Because of the natural constraint $\sum_{l=0}^{L} \pi_l = 1$, we treat π_0 as a reference category; this means that the probability for π_0 is determined by the remaining probabilities: $\pi_0 = 1 - \sum_{l=1}^{L} \pi_l$. The partial derivatives of the μ_{ij} with respect to L skill patterns could then result in the following $(L-1) \times J$ derivative matrix as follows:

$$\boldsymbol{D}_i = \frac{\partial \mu_{ij}}{\partial \pi_l} = \begin{bmatrix} \frac{\partial \mu_{i1}}{\partial \pi_1} & \cdots & \frac{\partial \mu_{iJ}}{\partial \pi_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_{i1}}{\partial \pi_L} & \cdots & \frac{\partial \mu_{iJ}}{\partial \pi_L} \end{bmatrix}. \tag{16}$$

Each element of the above matrix can be formulated as

$$\frac{\partial \mu_{ij}}{\partial \pi_{l'}} = \frac{\partial \sum_{l=0}^{L} \pi_l P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_l)}{\partial \pi_{l'}}, \tag{17}$$

where $l' = 1, \ldots, L$ ($L = 2^K - 1$). Each element of the partial derivatives with respect to the l'th pattern can be expressed as follows:

$$\begin{aligned} \frac{\partial \mu_{ij}}{\partial \pi_{l'}} &= \frac{\partial \sum_{l=0}^{L} \pi_l P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_l)}{\partial \pi_{l'}} \\ &= \frac{\partial \sum_{l=1}^{L} \pi_l P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_l) + \partial \pi_0 P(Y_{ij} = 1 | \boldsymbol{\alpha}_i = \boldsymbol{\alpha}_0)}{\partial \pi_{l'}} \\ &= \frac{\partial \sum_{l=1}^{L} \pi_l P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_l) + \partial \left(1 - \sum_{l=1}^{L} \pi_l\right) P(Y_{ij} = 1 | \boldsymbol{\alpha}_0)}{\partial \pi_{l'}} \\ &= P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_{l'}) - P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_0) \end{aligned}. \tag{18}$$

The first term in Eq. (18) is equivalent to 'anti' slip rate, $(1 - s_j)$, if skill pattern $\boldsymbol{\alpha}_{l'}$ is sufficient to answer item *j* correctly - i.e. $\eta_j(\boldsymbol{\alpha}_{l'}) = 1$; otherwise, if $\eta_j(\boldsymbol{\alpha}_{l'}) = 0$, it equals g_j. The second term in Eq. (18) is equal to g_j, because all the skills are absent from the pattern $\boldsymbol{\alpha}_0$. Thus Eq. (18) can be simplified as follows:

$$\frac{\partial \mu_{ij}}{\partial \pi_{l'}} = \begin{cases} (1 - s_j) - g_j & \text{if } \eta_j(\boldsymbol{\alpha}_{l'}) = 1 \\ 0 & \text{if } \eta_j(\boldsymbol{\alpha}_{l'}) = 0 \end{cases}. \tag{19}$$

In practice, some non-zero skill pattern(s) may not be sufficient to result in correct response(s) to any of the item(s) on a test. The term 'non-zero' indicates any pattern except the one in which all of the skills are absent: $\boldsymbol{\alpha}_0 = (0, \ldots, 0)^T$. Given the conjunctive rule of the DINA model formula, those skill patterns are essentially indistinguishable from the $\boldsymbol{\alpha}_0$ and also from each other. Thus, the probabilities of these skill patterns are not estimable. So, we constrain the probabilities of such classes to zero, and do not include them in the variance calculation. Note that this will lead to a corresponding reduction in rows of the derivative matrix. The above partial derivative matrix $\boldsymbol{D}_i$ in Eq. (16) is, thereby, incorporated to the bread and meat portions in Eqs. (6) and (7). With all the results from Sections 3.1, 3.2 and 3.4 incorporated to Eqs. (6) and (7), the sandwich covariance matrix Eq. (5) for $(L-1)$ skill pattern probabilities, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\pi}}_A)$ can be obtained.

Finally, in order to get the sandwich covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\pi})$ between the probabilities of *K* individual skills, the covariance matrix among the skill pattern probabilities should be transformed into the corresponding formula. Let $\boldsymbol{\Sigma}(\boldsymbol{\pi})$ denote a sandwich covariance matrix among probabilities of the *K* skills, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$. The $\boldsymbol{\Sigma}(\boldsymbol{\pi})$ can be obtained by using

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \mathbf{C}\boldsymbol{\Sigma}(\boldsymbol{\pi}_A)\mathbf{C}^T, \tag{20}$$

where $\mathbf{C}$ is a matrix that comprises all possible skill patterns, $K \times (L-1)$, except the $\boldsymbol{\alpha}_0$. For example, if $K = 2$, the skill pattern matrix and the covariance matrix among the skill pattern probabilities are as follows:

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}(\boldsymbol{\pi}_A) = \begin{bmatrix} \sigma_{[01,01]} & \sigma_{[01,10]} & \sigma_{[01,11]} \\ \sigma_{[10,01]} & \sigma_{[10,10]} & \sigma_{[10,11]} \\ \sigma_{[11,01]} & \sigma_{[11,10]} & \sigma_{[11,11]} \end{bmatrix}.$$

Therefore, the $(K \times K)$ covariance matrix among the probabilities of the K skills is

$$\boldsymbol{\Sigma}(\boldsymbol{\pi}) = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \boldsymbol{\Sigma}(\boldsymbol{\pi}_A) \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}^T.$$

Note that if there are non-zero skill pattern(s) that do not adequately provide correct response(s) to any of the item(s), there will be a corresponding reduction in the columns of $\mathbf{C}$, and the rows of $\boldsymbol{\Sigma}(\boldsymbol{\pi}_A)$. The square root of the diagonal elements of the estimated covariance matrix, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\pi}})$ results in the sandwich SEs of the skill probabilities.

4 Simulation study

The goal of the study is to evaluate the two proposed sandwich variance estimators of the DINA model parameters, through various simulation conditions in the presence of the cluster sampling design. We compare their performance with the jackknife estimator in regard of their accuracy and computational efficiency.

4.1 Study design

Because our interest lies in the clustered data, the number of clusters was set to T = 15, 30, 60 and 90; and the cluster sizes were n_T = 25, 50 and 75, those being the number of examinees included in each cluster. In the data-generation step, we employed a Q-matrix that contains 35 items for measuring five skills in total (Table 1). The matrix was originally used in Ravand et al. (2013) for a reading comprehension test. In order to manipulate various realistic situations of item quality, 35 guessing parameters were randomly generated from the uniform distribution, Unif(0, 0.3); then the corresponding (1 − slip) parameters were generated from the uniform distribution, Unif(guess + 0.3, 1); the guess and slip values are also listed in Table 1.

Table 1 Q-matrix and generating item parameters in the simulation design

Item No.	α_1	α_2	α_3	α_4	α_5	g_j	s_j
1	0	0	0	1	0	0.169	0.430
2	0	0	1	0	0	0.218	0.268
3	0	0	1	0	0	0.275	0.144
4	0	0	1	0	0	0.009	0.333

Table 1 Q-matrix and generating item parameters in the simulation design (continued)

Item No.	α_1	α_2	α_3	α_4	α_5	g_j	s_j
5	0	1	1	0	0	0.230	0.075
6	0	0	0	0	1	0.144	0.437
7	1	0	0	0	0	0.027	0.453
8	0	0	1	0	0	0.259	0.121
9	0	0	0	1	1	0.149	0.096
10	0	0	0	1	1	0.256	0.373
11	1	0	0	1	0	0.188	0.062
12	0	0	0	1	0	0.072	0.144
13	0	0	0	1	1	0.011	0.629
14	1	0	0	0	0	0.155	0.393
15	1	0	0	0	0	0.246	0.362
16	0	0	0	1	1	0.149	0.305
17	0	0	1	1	0	0.054	0.023
18	0	1	1	0	0	0.264	0.149
19	0	0	1	0	0	0.131	0.361
20	0	0	0	1	1	0.218	0.197
21	0	0	1	0	0	0.259	0.379
22	0	0	1	0	0	0.090	0.549
23	0	0	0	1	0	0.282	0.203
24	1	0	0	0	0	0.128	0.172
25	0	0	1	0	0	0.106	0.402
26	1	0	1	1	0	0.126	0.244
27	1	0	0	0	0	0.093	0.090
28	0	0	0	1	1	0.021	0.442
29	0	0	0	1	0	0.088	0.275
30	0	1	0	0	0	0.241	0.333
31	0	1	0	0	0	0.212	0.100
32	1	0	1	0	0	0.292	0.134
33	0	1	0	0	0	0.022	0.045
34	0	1	1	0	0	0.300	0.295
35	0	0	0	1	0	0.164	0.286

In simulating clustering effects, we first induced positive correlations ($=\boldsymbol{\Sigma}(\boldsymbol{\pi})$) between the five skill probabilities ($=5 \times 5$), which are varied across clusters. We generated true values for the moderate skill probabilities (i.e. $\pi_1,\ldots,\pi_5$) based on the beta distribution, *Beta* (4, 8), with a mean of 0.33 and a standard deviation of 0.13 for each π_k. We chose the beta distribution because the skill probability has values in the (0, 1) interval, and is the conjugate prior distribution of the binomial distribution (Patterson et al., 2002).

If $T = 3$ clusters, e.g. the true skill probabilities $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ for each cluster appear to be (0.48, 0.34, 0.41, 0.49, 0.43) for cluster 1, (0.30, 0.42, 0.30, 0.14, 0.49) for cluster 2 and (0.29, 0.25, 0.34, 0.46, 0.26) for cluster 3. To generate examinee-level sampling weights, we used a gamma distribution with 15 as the shape parameter and 30 as the scale parameter in each simulation condition. The distribution mimics the sampling distribution of the TIMSS assessment for grade 8 students in the USA. For each of the 12 simulation conditions as combinations of the $T = 15$, 30, 60 and 90 (clusters) and $n_T = 25$, 50 and 75 (examinees), we generated a total of 100 independent data sets.

In each condition, the expectation-maximisation (EM) algorithm was implemented by using the R package 'CDM' (Robitzsch et al., 2014). With the convergence criterion of 10^{-5} as the maximal change in parameter estimates and maximum number of 5,000 iterations, the model parameters were estimated separately for the 100 replications of the data sets in each condition. The R code for the standard error estimation procedures within R 3.0.1 is provided in Appendix A.

In computing the jackknife SE, the following steps were taken. Let $\hat{\beta}$ denote the estimator based on the entire cluster and $\hat{\beta}_{(-t)}$ denote the estimator for the data excluding the cluster *t*. Then the jackknife variance estimator can be formulated based upon:

$$\Sigma(\hat{\beta})_{JK} = \frac{(T-1)}{T}\sum_{t=1}^{T}\left(\hat{\beta}_{(-t)} - \overline{\beta}_{JK}\right)^2,$$

where $\overline{\beta}_{JK} = \frac{1}{T}\sum_{t=1}^{T}\hat{\beta}_{(-t)}$. Specifically, the T ($t = 1,\ldots,T$) estimators for each of the g_j ($j = 1,\ldots,J$), s_j ($j = 1,\ldots,J$) and π_l ($l = 1,\ldots,L$) can be computed, with a different cluster being deleted, in each case. Finally, the jackknife SE can then be obtained by taking square roots of the diagonal elements of the $\Sigma(\hat{\beta})_{JK}$.

The results of the simulation study were summarised in several aspects. Let $\mathrm{SE}(\hat{g}_{bj})$, $\mathrm{SE}(\hat{s}_{bj})$ and $\mathrm{SE}(\hat{\pi}_{bk})$ indicate the resulting SE estimates for each item *j* or skill *k*, corresponding to the replication *b*. First, in order to compare the computational efficiency of the jackknife and sandwich estimators, the elapsed central processing unit (CPU) times were measured in each condition. The elapsed time includes the time taken to implement EM algorithm, and then to calculate the SEs. Next, to evaluate the accuracy of the three estimators, we obtained the criterion (or true) SEs by empirically calculating the standard deviations of the point estimates for all 100 of the replicated data sets as follows:

- $\mathrm{ESD}(g_j) = \frac{1}{(100-1)}\sum_{b=1}^{100}(\hat{g}_{bj} - g_{bj})^2$ for guess
- $\mathrm{ESD}(s_j) = \frac{1}{(100-1)}\sum_{b=1}^{100}(\hat{s}_{bj} - s_{bj})^2$ for slip
- $\mathrm{ESD}(\pi_k) = \frac{1}{(100-1)}\sum_{b=1}^{100}(\hat{\pi}_{bk} - \pi_{bk})^2$ for skill probability

where $j = 1,\ldots,35$ (items) and $k = 1,\ldots,5$ (skills), and $b = 1,\ldots,100$ (replications). Using the empirical standard deviation (ESD) as the criterion, the performance of each of

the SE estimators was evaluated with regard to relative error and absolute bias. The relative error was used to display the direction and relative distance of the estimated SEs to the corresponding ESDs as follows:

- Relative error $(\mathrm{SE}(\hat{g}_{bj})) = \dfrac{1}{\mathrm{ESD}(g_j)} \{\mathrm{SE}(\hat{g}_{bj}) - \mathrm{ESD}(g_j)\}$ for guess
- Relative error $(\mathrm{SE}(\hat{s}_{bj})) = \dfrac{1}{\mathrm{ESD}(s_j)} \{\mathrm{SE}(\hat{s}_{bj}) - \mathrm{ESD}(s_j)\}$ for slip
- Relative error $(\mathrm{SE}(\hat{\pi}_{bk})) = \dfrac{1}{\mathrm{ESD}(\pi_k)} \{\mathrm{SE}(\hat{\pi}_{bk}) - \mathrm{ESD}(\pi_k)\}$ for skill probability

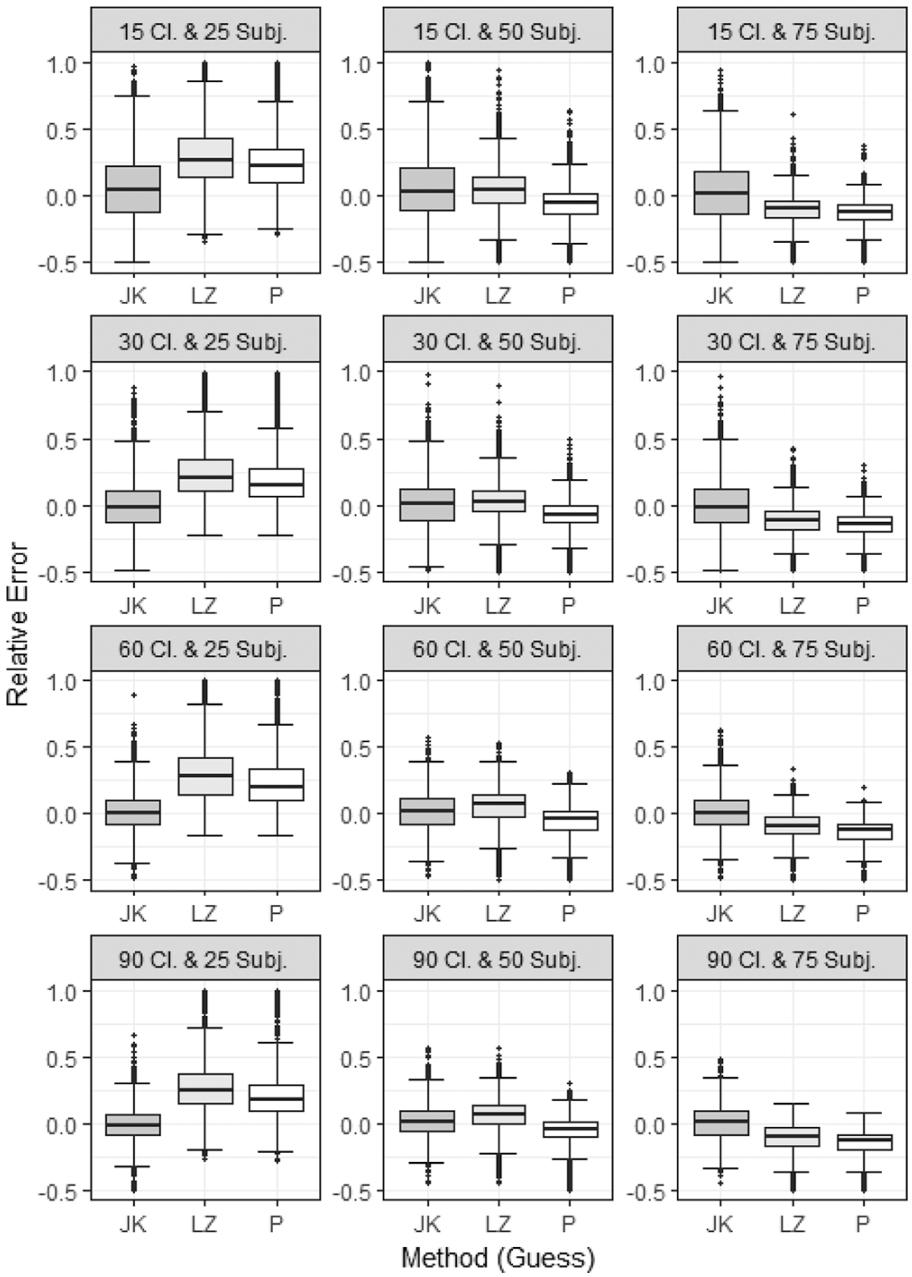
The absolute bias was used to calculate the bias of the absolute distance of the estimated SEs from the corresponding ESDs, across the 100 replications and the 35 items (or five skills), as follows:

- $|\mathrm{Bias}(\mathrm{SE}(\hat{g}_j))| = \dfrac{1}{35}\sum_{j=1}^{35}\left\{\dfrac{1}{100}\sum_{b=1}^{100}|\mathrm{SE}(\hat{g}_{bj}) - \mathrm{ESD}(g_j)|\right\}$ for guess
- $|\mathrm{Bias}(\mathrm{SE}(\hat{s}_j))| = \dfrac{1}{35}\sum_{j=1}^{35}\left\{\dfrac{1}{100}\sum_{b=1}^{100}|\mathrm{SE}(\hat{s}_{bj}) - \mathrm{ESD}(s_j)|\right\}$ for slip
- $|\mathrm{Bias}(\mathrm{SE}(\hat{\pi}_k))| = \dfrac{1}{5}\sum_{k=1}^{5}\left\{\dfrac{1}{100}\sum_{b=1}^{100}|\mathrm{SE}(\hat{\pi}_{bk}) - \mathrm{ESD}(\pi_k)|\right\}$ for skill probability

4.2 Results

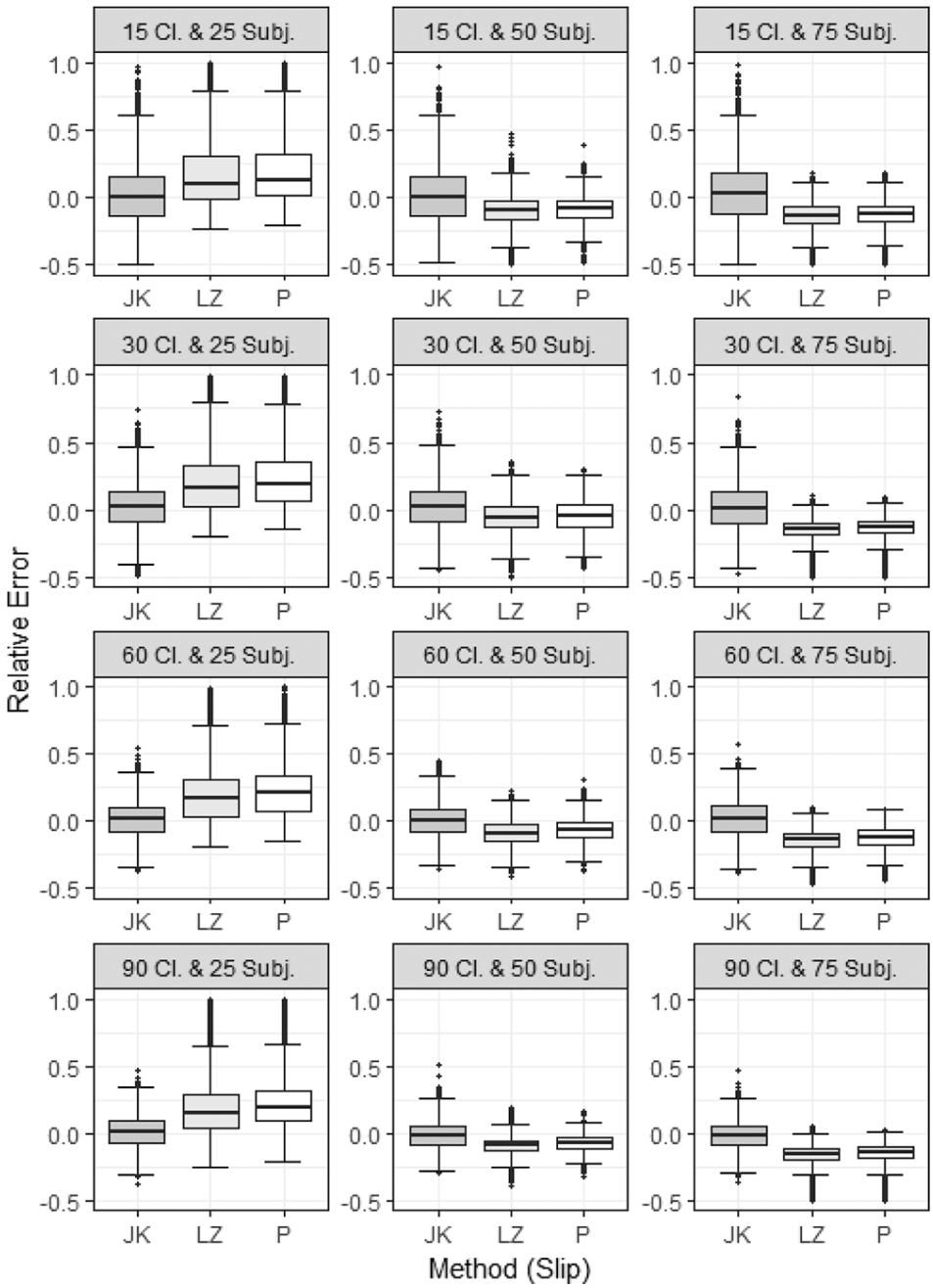
Figures 1–3 contain boxplots demonstrating the performances of the JK, Liang and Zeger (LZ) and Pan-modified SEs for the DINA model parameters. More specifically, each figure displays the relative errors corresponding to guess parameters (Figure 1), slip parameters (Figure 2) and skill probabilities (Figure 3), respectively. Each panel in the figures represents a different simulation condition with a varying number of clusters (T = 15, 30, 60 and 90) and cluster sizes (n_T = 25, 50 and 75), with the jackknife and the two sandwich estimators being on the x-axis and the relative error on the y-axis. The boxplots in each panel comprise the relative errors across 100 simulated data sets and across all 35 guess/slip parameters and five skills. In other words, each boxplot in Figures 1 and 2 was constructed by 3,500 (=100 replications × 35 guess/slip parameters) outcomes; each boxplot in Figure 3 was constructed by 500 (=100 replications × 5 skills) outcomes. Overall, the results seen in Figures 1–3 describe the effects of the number of clusters and the cluster size.

Figure 1 Relative errors for guess parameters



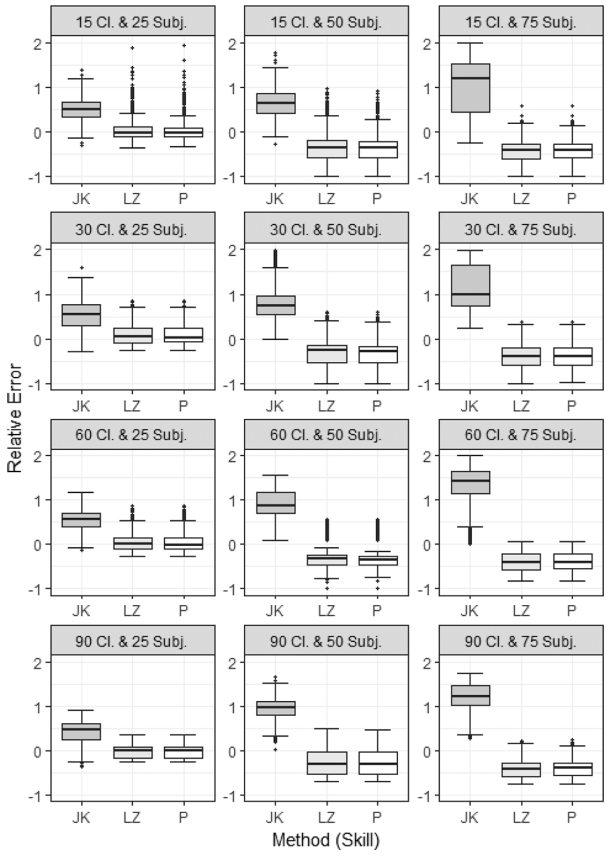
Note: Cl, number of clusters (T); JK, jackknife; LZ, Liang and Zeger; P, Pan-modified; Subj., cluster size (n_T)

Figure 2 Relative errors for slip parameters



Note: Cl, number of clusters (T); JK, jackknife; LZ, Liang and Zeger; P, Pan-modified; Subj., cluster size (n_T)

Figure 3 Relative errors for skill probabilities



Note: Cl, number of clusters (T); JK, jackknife; LZ, Liang and Zeger; P, Pan-modified; Subj., cluster size (n_T)

4.2.1 Comparison of relative error

Figures 1 and 2 for the guess and slip parameters showed that the LZ and Pan-modified estimators tend to be sensitive to small cluster size ($n_T = 25$). Specifically, they revealed upward relative errors and wide dispersion in estimating the corresponding SEs. When it comes, however, to the moderate ($n_T = 50$) to large cluster sizes ($n_T = 75$), the dispersions for the both estimators rapidly diminished. The relative errors from the JK were centred around zero for the guessing and slip parameters in any condition; but the dispersion tended not to be diminished across different conditions.

As shown in Figure 3 for the skill probability, overall, it is remarkable that the sandwich estimators (i.e. LZ and Pan-modified) revealed smaller errors than the JK, regardless of the numbers of clusters and cluster sizes. The LZ and Pan-modified estimators were affected by cluster sizes; they led to tiny upward errors for the small cluster size ($n_T = 25$), but the errors got smaller with moderate to large cluster sizes ($n_T = 50$ or 75). On the other hand, the JK estimator typically produced considerable errors as the number of clusters and cluster sizes increased. Also, the magnitude of error for the JK had severe increases for the larger cluster sizes.

4.2.2 Computational time

Table 2 (see the fourth column) shows the results pertaining to the summaries of the elapsed CPU time in each simulation condition. Note that for the JK the computational time ranged from 709 to 47,595 s, for the LZ from 333 to 4,813 s and for the Pan-modified from 399 to 6,498 s.

Table 2 Simulation results of guess, slip and skill probability

				Guess		*Slip*		*Skill*	
T	n_T	*Method*	*Elapsed (s)*	*ASE*	$\|Bias\| \times 10$	*ASE*	$\|Bias\| \times 10$	*ASE*	$\|Bias\| \times 10$
15	25	JK	709	0.033	0.071	0.029	0.051	0.040	0.134
		LZ	333	0.046	0.146	0.033	0.062	0.030	0.064
		P	399	0.044	0.126	0.034	0.063	0.030	0.064
	50	JK	1,107	0.017	0.024	0.014	0.019	0.026	0.114
		LZ	644	0.017	0.017	0.013	0.014	0.011	0.052
		P	670	0.015	0.018	0.013	0.013	0.011	0.053
	75	JK	1,659	0.010	0.010	0.008	0.009	0.018	0.091
		LZ	892	0.009	0.012	0.007	0.012	0.006	0.029
		P	1,010	0.008	0.015	0.007	0.011	0.006	0.028
30	25	JK	2,187	0.024	0.047	0.020	0.034	0.033	0.128
		LZ	593	0.024	0.028	0.018	0.024	0.015	0.093
		P	672	0.021	0.026	0.018	0.021	0.015	0.094
	50	JK	3,652	0.014	0.020	0.012	0.015	0.025	0.134
		LZ	1,228	0.012	0.020	0.010	0.017	0.008	0.041
		P	1,336	0.012	0.023	0.010	0.016	0.008	0.040
	75	JK	5,510	0.014	0.013	0.011	0.011	0.016	0.047
		LZ	1,633	0.019	0.054	0.013	0.025	0.013	0.016
		P	2,004	0.018	0.044	0.014	0.028	0.013	0.016
60	25	JK	7,749	0.019	0.035	0.016	0.027	0.031	0.149
		LZ	1,017	0.017	0.025	0.014	0.023	0.011	0.063
		P	1,315	0.016	0.027	0.014	0.022	0.011	0.061
	50	JK	16,528	0.017	0.019	0.014	0.014	0.021	0.068
		LZ	2,068	0.023	0.067	0.017	0.031	0.016	0.030
		P	2,640	0.022	0.056	0.017	0.034	0.016	0.031
	75	JK	33,574	0.010	0.009	0.008	0.006	0.015	0.069
		LZ	3,061	0.010	0.009	0.007	0.008	0.007	0.027
		P	3,992	0.009	0.008	0.008	0.007	0.007	0.028
90	25	JK	16,522	0.023	0.035	0.020	0.026	0.027	0.088
		LZ	1,496	0.032	0.087	0.024	0.045	0.022	0.040
		P	2,123	0.030	0.072	0.024	0.048	0.022	0.039

Table 2 Simulation results of guess, slip and skill probability (continued)

				Guess		*Slip*		*Skill*	
T	n_T	*Method*	*Elapsed (s)*	*ASE*	*|Bias| × 10*	*ASE*	*|Bias| × 10*	*ASE*	*|Bias| × 10*
	50	JK	35,141	0.012	0.013	0.010	0.010	0.018	0.081
		LZ	3,061	0.012	0.013	0.009	0.011	0.008	0.036
		P	4,741	0.011	0.011	0.009	0.009	0.008	0.037
		JK	47,595	0.008	0.008	0.007	0.006	0.014	0.073
	75	LZ	4,813	0.007	0.010	0.006	0.011	0.005	0.024
		P	6,498	0.007	0.012	0.006	0.010	0.005	0.024

Note: ASE, averaged standard error; JK, jackknife; LZ, Liang and Zeger; n_T, sample size per cluster; P, Pan-modified; T, number of clusters.

4.2.3 Comparison of absolute bias

Table 2 also shows the average estimated SE (=ASE), and the average absolute bias (=|Bias|) of the estimators in each simulation condition. The numerical results for the |Bias| were rescaled by multiplying by 10, such that |Bias| × 10. The absolute biases serve to confirm the relative error shown in Figures 1–3. The magnitudes of the biases from all estimators decreased as number of clusters and cluster sizes increased. Overall, the sandwich SEs approximated the ESD as accurately as the JK SE for guess and slip parameters. In contrast, the sandwich estimators outperformed with the greater accuracy as for skill probability, specifically when the number of clusters increased by moderate to large cluster sizes (n_T =50 or 75); occasionally, the absolute bias from the LZ and Pan-modified estimators were two to three times smaller than that from the JK estimator.

5 Real data example: TIMSS 2011 mathematics

We applied the proposed variance estimators to a real data example: mathematics test in the TIMSS 2011. Current analysis is based on grade 8 test data drawn from the benchmark population in the USA. It comprises a total of nine states: Alabama, California, Colorado, Connecticut, Indiana, Florida, Massachusetts, Minnesota and North Carolina. Information about 89 of the mathematics test items in the TIMSS was released, and it can be found on the database's website (http://timssandpirls.bc.edu/timss2011/international-released-items.html). Noting the exception of those examinees who were not given any of the 89 items, the sample size from the benchmark states was n = 11,158 (examinees).

For this study all items were scored dichotomously, with the highest possible score being 1 and all the others being 0. A Q matrix for the 89 released items was developed by two mathematics educators using the National Council of Teachers of Mathematics (2000) and the TIMSS 2011 frameworks (Mullis et al., 2009). Four content domains - number and operation, algebra, geometry and data and probability - were used to identify topic areas evaluating students' understanding of mathematics, and this approach resulted in a total of nine attributes. In other words, nine content-based attributes were defined as fine-grained skills needed to solve the 89 released items. The complete list of skills and the Q-matrix are given in Appendix B.

The TIMSS uses a stratified two-stage cluster-sampling design (Joncas and Foy, 2012). For example, schools were primarily sampled within each state, and classrooms were sampled within each school. The numbers of schools within the nine states are 55, 53, 62, 82, 60, 56, 56, 54 and 59. TIMSS manual notes that the schools within each state were paired to construct sampling zones for the purpose of using the jackknife method to calculate the sampling structure; thus the numbers of sampling zones are 28, 27, 32, 42, 30, 28, 28, 27 and 30. Within each sampling zone, one of two schools was randomly selected to have its contribution doubled and the other one to have zero contribution. The database also provides the examinee-level sampling weight within each state. That is given as $w_i = w_{\text{sc}} \times w_{\text{cl}} \times w_{\text{st}}$, where w_{sc} is the school-level weight with a school nonparticipation adjustment, w_{cl} is the basic class-level weight for all sampled classes in each school with a class nonparticipation adjustment and w_{st} is the examinee-level weight for each examinee in the classroom of the school with a student nonparticipation adjustment (Joncas and Foy, 2012). In Eq. (10) of the Pan-modified estimator, we estimated the true variance-correlation matrix by pooling all of the examinees within each state. The working correlation matrices for both LZ and Pan-modified estimators considered different correlation structures across states.

Table 3 showed the results for guess and slip parameters corresponding to 23 randomly selected items out of the 89 in total. The table comprises skill patterns required to correctly answer those items, guess and slip parameter estimates, with the corresponding SE estimates obtained from the jackknife and sandwich estimators. Across all 89 items the averaged SEs for the guess parameters were 0.022, 0.016 and 0.025 for the JK, LZ and Pan, respectively. The averaged SEs for the slip parameters were 0.021, 0.014 and 0.019, respectively. On average, the jackknife resulted in biggest and the LZ resulted in the smallest SEs.

Table 3 TIMSS 2011 assessments for nine benchmark states in US: parameter estimates and standard errors for the guess and slip parameters for 23 out of 89 selected items

		Guess					*Slip*				
			SE					*SE*			
Item	*Q-matrix*	*Est.*	*Naïve*	*JK*	*LZ*	*P*	*Est.*	*Naïve*	*JK*	*LZ*	*P*
1	001000010	0.332	0.012	0.019	0.016	0.015	0.447	0.018	0.025	0.022	0.021
2	101000010	0.031	0.004	0.008	0.005	0.006	0.300	0.018	0.029	0.023	0.021
3	100000000	0.443	0.015	0.031	0.023	0.030	0.143	0.011	0.016	0.012	0.018
4	010000010	0.301	0.012	0.026	0.018	0.034	0.049	0.006	0.017	0.006	0.019
5	000101000	0.061	0.006	0.008	0.009	0.007	0.811	0.014	0.015	0.015	0.013
6	100100000	0.002	0.001	0.002	0.001	0.001	0.798	0.014	0.021	0.014	0.009
7	100001000	0.273	0.011	0.022	0.018	0.020	0.150	0.012	0.019	0.013	0.016
8	000001100	0.099	0.008	0.016	0.009	0.009	0.262	0.012	0.031	0.016	0.017
9	101000000	0.433	0.013	0.023	0.020	0.026	0.089	0.010	0.026	0.010	0.020
10	100100000	0.586	0.013	0.029	0.020	0.036	0.060	0.009	0.011	0.009	0.019
11	101000000	0.212	0.011	0.021	0.016	0.013	0.308	0.017	0.034	0.020	0.017
12	100100000	0.181	0.009	0.020	0.016	0.013	0.296	0.018	0.025	0.019	0.016

Table 3 TIMSS 2011 assessments for nine benchmark states in US: parameter estimates and standard errors for the guess and slip parameters for 23 out of 89 selected items (continued)

		Guess					*Slip*				
			SE					*SE*			
Item	*Q-matrix*	*Est.*	*Naïve*	*JK*	*LZ*	*P*	*Est.*	*Naïve*	*JK*	*LZ*	*P*
13	100001000	0.359	0.013	0.026	0.021	0.030	0.063	0.008	0.013	0.009	0.016
14	010000010	0.546	0.014	0.029	0.022	0.025	0.186	0.013	0.024	0.018	0.020
15	100100000	0.378	0.012	0.023	0.020	0.029	0.070	0.010	0.019	0.007	0.017
16	010010000	0.184	0.011	0.016	0.015	0.012	0.452	0.017	0.028	0.023	0.018
17	100001000	0.229	0.010	0.022	0.015	0.017	0.205	0.014	0.028	0.015	0.018
18	010000010	0.305	0.012	0.023	0.018	0.021	0.167	0.013	0.019	0.013	0.019
19	100000010	0.502	0.013	0.022	0.020	0.025	0.183	0.014	0.015	0.014	0.020
20	100100000	0.406	0.012	0.028	0.018	0.029	0.085	0.010	0.014	0.009	0.018
21	000101000	0.301	0.011	0.019	0.018	0.020	0.280	0.017	0.026	0.018	0.025
22	000000110	0.335	0.013	0.023	0.018	0.028	0.008	0.001	0.009	0.002	0.009
23	000010110	0.097	0.007	0.019	0.009	0.013	0.056	0.004	0.018	0.005	0.014
	Average	0.329	0.010	0.022	0.016	0.025	0.255	0.012	0.021	0.014	0.019

Note: Est., estimated guess and slip parameters ($\hat{g}_j$ and $\hat{s}_j$); JK, jackknife; LZ, Liang and Zeger's estimator; P, modified Pan's estimator.

Table 4 shows the results for the nine skill mastery probabilities along with the estimated SEs by using the three methods. The estimated skill probabilities range from 0.385 (pattern) to 0.584 (data organisation, representation and interpretations). In other words, the eighth graders in the US benchmark states tend to be least proficient in *pattern*, within the domain of algebra, and most proficient in *data organisation, representation* and *interpretations*, within the domain of data and probability. Overall, the results show that more than 50% of the eighth graders were proficient in the skills related to the data and probability domain (data organisation, representation and interpretation; probability), and in the skills related to whole number and integer within the number domain. Notice that the JK and Pan estimators performed more similarly than the LZ estimator. The LZ tends to be consistently smaller than those produced by other estimators. Across all nine skills the averaged SEs for the skill probabilities were 0.018, 0.007 and 0.017 for the JK, LZ and Pan, respectively.

Table 4 TIMSS 2011 assessments for nine benchmark states in US: parameter estimates and standard errors of nine skill probabilities

				Sandwich	
Skill		*Est.*	*JK*	*LZ*	*P*
1	Whole numbers and integers	0.533	0.014	0.011	0.017
2	Fractions, decimals and proportions	0.461	0.015	0.005	0.014
3	Patterns	0.385	0.021	0.002	0.005

Table 4 TIMSS 2011 assessments for nine benchmark states in US: parameter estimates and standard errors of nine skill probabilities (continued)

			Sandwich	
Skill	*Est.*	*JK*	*LZ*	*P*
4 Expressions, equations and functions	0.492	0.015	0.005	0.034
5 Lines, angles and shapes	0.432	0.019	0.005	0.007
6 Measurement	0.496	0.020	0.012	0.016
7 Location and movement	0.392	0.014	0.011	0.028
8 Data organisation, representation and interpretations	0.584	0.022	0.008	0.014
9 Probability	0.512	0.021	0.007	0.018
Average	0.476	0.018	0.007	0.017

Note: Est., estimated skill probability ($\hat{\pi}_k$); JK, jackknife; LZ, Liang and Zeger's estimator; P, modified Pan's estimator.

Computational time revealed some dramatic differences between the JK and the sandwich estimators. The procedure took approximately 4 h 7 min using the JK method, whereas 2 min when using each of the sandwich methods.

6 General diagnostic model

There are occasions when use of the more generalised family of the CDMs such as GDM (von Davier, 2005), Loglinear Cognitive Diagnosis Model (LCDM) (Henson et al., 2009), or generalised DINA (de la Torre, 2011) are preferred. The proposed sandwich formulation in Section 3 can readily be applied to the above-mentioned models although partial derivatives have to be recalculated (see Sections 3.3 and 3.4) because of different parameterisations of the models.

Let us use the GDM as an example of the generalised models. Suppose both the response outcome and attribute proficiency are dichotomous. Using a logistic link function, the model can be formulated as:

$$P(\boldsymbol{Y}_{ij}=1|\boldsymbol{\alpha}_l)=\frac{\exp\left[\beta_j+\gamma_j^T h(\boldsymbol{q}_j,\boldsymbol{\alpha})\right]}{1+\exp\left[\beta_j+\gamma_j^T h(\boldsymbol{q}_j,\boldsymbol{\alpha})\right]}, \tag{21}$$

where $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_K)$ is the attribute proficiency; $\boldsymbol{q}_j$ is the set of skills influencing item j as given by the jth row of the Q-matrix, β_j is intercept parameter and γ_j is slope parameter. For this example, we use $h(\boldsymbol{q}_j,\boldsymbol{\alpha})$ is formulated as - $h(\boldsymbol{q}_j,\boldsymbol{\alpha})=(q_{j1}\alpha_1,\ldots,q_{jK}\alpha_K)$, i.e. a fully saturated model - so Eq. (21) becomes $P(\boldsymbol{Y}_{ij}=1|\boldsymbol{\alpha}_l)=\dfrac{\exp\left[\beta_j+\sum_k\gamma_{jk}q_{jk}\alpha_k\right]}{1+\exp\left[\beta_j+\sum_k\gamma_{jk}q_{jk}\alpha_k\right]}$.

As shown in Eq. (2), the marginal probability for answering item j correctly is $\mu_{ij} = P(\boldsymbol{Y}_{ij} = 1) = \sum_{l=0}^{L} \pi_l P(\boldsymbol{Y}_{ij} = 1|\boldsymbol{\alpha}_l)$. Similar to Sections 3.3 and 3.4 for the DINA model, the mathematical calculation of derivative matrices for the GDM parameters in Eq. (21) is necessary. For μ_{ij} , each element of the partial derivative matrix, with respect to β_j and γ_{jk} , is simplified as follows:

$$\frac{\partial \mu_{ij}}{\partial \beta_{j'}} = \sum_{l=1}^{L} \frac{\exp\left[\beta_j + \sum_k \gamma_{jk} q_{jk} \alpha_k\right]}{\{1 + \exp\left[\beta_j + \sum_k \gamma_{jk} q_{jk} \alpha_k\right]\}^2} \pi_l \quad \text{if } j = j' = 0, \text{ otherwise.}$$

$$\frac{\partial \mu_{ij}}{\partial \gamma_{j'k}} = \sum_{l=1}^{L} \frac{\exp\left[\beta_j + \sum_k \gamma_{jk} q_{jk} \alpha_k\right]}{\{1 + \exp\left[\beta_j + \sum_k \gamma_{jk} q_{jk} \alpha_k\right]\}^2} q_{jk} \alpha_k \pi_l \quad \text{for item } j = j' \text{ and}$$
$$k = 1, \ldots, K = 0, \text{ otherwise.}$$

Finally, each element of the partial derivative with respect to the skill probability, $\frac{\partial \mu_{ij}}{\partial \pi_l}$, can be shown to be, using the same logic as the DINA model in Section 3.4:

$$\frac{\partial \mu_{ij}}{\partial \pi_{l'}} = P(Y_{ij} = 1|\boldsymbol{\alpha}_{l'}) - P(Y_{ij} = 1|\boldsymbol{\alpha}_0).$$

7 Discussion

This study has shown how we developed accurate and faster methods to calculate the standard errors associated with the DINA model. We mainly focused on the scenarios in which cluster sampling design is embedded. The sandwich variance formula for two item parameters, guess and slip, and latent skill probability in the model were derived. Two approaches to formulating the sandwich variance were examined. We used the approach originally proposed by Liang and Zeger (1986) and adopted and modified the approach proposed by Pan (2001). The key difference between the two sandwich estimators pertains to estimating the true variance–covariance matrix between item responses. The variance-covariance matrix is estimated by each examinee individually in the former approach, by pulling all examinees within each cluster in the latter approach. Finally, we evaluated the performance by applying the jackknife replication technique, to combinations of various numbers of clusters and cluster sizes.

The proposed sandwich estimator

a accurately took into account the clustered structure of the data

b was much faster than the resampling techniques when dealing with large scale data

c was straightforward to code in the existing statistical program.

The simulation results have several implications with regard to the estimators. First, the LZ estimator consistently revealed weaknesses in handling small sample size per cluster. That poor performance suggests that it is not an efficient tool for use in estimating the variance–covariance matrix with the only responses from a single examinee. In the

presence of the small sample size per cluster, the inefficiency increases along with the increasing number of clusters. Second, the jackknife technique exposed a general limitation with respect to the latent skill probabilities. The jackknife estimator tended to overestimate the true standard errors regardless of any condition. In fact, our finding supports Patterson et al. (2002), within that the technique tends to overestimate the standard errors of the latent class probabilities in the traditional LCM. The discrepancy got considerably bigger with a large sample size per cluster. Thus, researchers should be cautious when they use the jackknife technique to conduct their empirical data analysis. Finally, with respect to the computational time, we found that the sandwich estimators were considerably more efficient than the jackknife.

Despite the fact that the Pan-modified estimator did a good job for adjusting for small sample sizes, we cannot ignore a number of studies that developed bias-corrected sandwich estimators in an attempt to alleviating bias due to the small sample size. Examples include Mancl and DeRouen (2001), Wang and Long (2011) and Li and Redden (2015), all of whose findings were formulated on the basis of the GEE. Therefore, some sort of comparison or even amalgamation of those studies with the current study might teach us much about the overall efficiency. Finally, this study has focused mainly on the DINA model. Depending upon the structure of the data, however, there are occasions when use of the more generalised family of the CDMs (de la Torre, 2011; Henson et al., 2009; von Davier, 2005, 2014) would be preferred over the DINA.

References

Betensky, R.A., Talcott, J.A. and Weeks, J.C. (2000) 'Binary data with two non-nested sources of clustering: an analysis of physician recommendations for early prostate cancer treatment', *Biostatistics*, Vol. 1, No. 2, pp.219–230.

Chen, H. and Chen, J. (2015) 'Exploring reading comprehension skill relationships through the G-DINA model', *Educational Psychology: An International Journal of Experimental Educational Psychology*, Vol. 36, No. 6, pp.1049–1064. doi:10.1080/01443410.2015.1076764.

Chen, J. and de la Torre, J. (2014) 'A procedure for diagnostically modeling extant large-scale assessment data: the case of the programme for international student assessment in reading', *Psychology*, Vol. 5, No. 18, pp.1967–1978.

Dayton, C.M. and Macready, G.B. (1988) 'Concomitant-variable latent class models', *Journal of the American Statistical Association*, Vol. 83, No. 401, pp.173–178.

de la Torre, J. (2011) 'The generalized DINA model framework', *Psychometrika*, Vol. 76, No. 2, pp.179–199.

de la Torre, J. and Lee, Y-S. (2013) 'Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis', *Journal of Educational Measurement*, Vol. 50, No. 4, pp.355–373.

de Leeuw, E.D., *Hox,* J.J. and *Dillman,* D.A. (Eds.) (2008) *The International Handbook of Survey Methodology*, Erlbaum/Taylor & Francis, New York/London.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004) *Applied Longitudinal Analysis*, Wiley, New York.

Haertel, E.H. (1989) 'Using restricted latent class models to map the skill structure of achievement items', *Journal of Educational Measurement*, Vol. 26, No. 4, pp.333–352.

Henson, R., Templin, J. and Willse, J. (2009) 'Defining a family of cognitive diagnosis models using log-linear models with latent variables', *Psychometrika*, Vol. 74, No. 2, pp.191–210.

Hou, L., de la Torre, J. and Nandakumar, R. (2014) 'Differential item functioning assessment in cognitive diagnostic modeling: application of the Wald test to investigate DIF in the DINA model', *Journal of Educational Measurement*, Vol. 51, No. 1, pp.98–125.

Hsieh, C., Xu, X. and von Davier, M. (2010) Variance Estimation for NAEP Data Using a Resampling-Based Approach: An Application of Cognitive Diagnostic Models, ETS Research Report Series, RR-10-26.

Huber, P.J. (1967) 'The behaviour of maximum likelihood estimates under non-standard conditions', in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, Berkeley, CA, pp.221–233.

Huber, P.J. (1981) *Robust Statistics*, Wiley, New York.

Ip, E.H. (2000) 'Adjusting for information inflation due to local dependency in moderately large item clusters', *Psychometrika*, Vol. 65, No. 1, pp.73–91.

Ip, E.H. (2010) 'Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models', *British Journal of Mathematical and Statistical Psychology*, Vol. 63, pp.395–416.

Ip, E.H. and Chen, S.H. (2012) 'Projective item response model for test-independent measurement', *Applied Psychological Measurement*, Vol. 36, No. 7, pp.581–601.

Johnson, E.G., Rust, K.F. and Hansen, M.H. (1990) 'Weighting procedures and estimation of sampling variance', in Johnson, E.G. and Zwick, R. (Eds.): *Focusing the New Design: The NAEP 1988 Technical Report* (No. 19-TR-20), Educational Testing Service, National Assessment of Educational Progress, Princeton, NJ.

Johnson, M.S., Lee, Y-S., Park, J.Y., Zhang, Z. and Sachdeva, R. (2013) 'Comparing attribute distribution across countries: application to TIMSS 2007 mathematics', *Presented at the annual meeting of the National Council on Measurement in Education*, San Francisco, CA.

Joncas, M. and Foy, P. (2012) 'Sample design in TIMSS and PIRLS', in Martin, M.O. and Mullis, I.V.S. (Eds.): *Methods and Procedures in TIMSS and PIRLS 2011*, TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

Junker, B.W. and Sijtsma, K. (2001) 'Cognitive assessment models with few assumptions, and connections with nonparametric item response theory', *Applied Psychological Measurement*, Vol. 25, No. 3, pp.258–272.

Lee, Y-S., Park, Y.S. and Taylan, D. (2011) 'A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007', *International Journal of Testing*, Vol. 11, pp.144–177.

Li, P. and Redden, D.T. (2015) 'Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes', *Statistics in Medicine*, Vol. 34, No. 2, pp.281–296.

Liang, K-Y. and Zeger, S.L. (1986) 'Longitudinal data analysis using generalized linear models', *Biometrika*, Vol. 73, No. 1, pp.13–22.

Mancl, L.A. and DeRouen, T.A. (2001) 'A covariance estimator for GEE with improved small-sample properties', *Biometrics*, Vol. 57, No. 1, pp.126–134.

Mullis, I.V.S., Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y. and Preuschoff, C. (2009) *TIMSS 2011 Assessment Frameworks*, TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA.

National Council of Teachers of Mathematics (2000) *Principles and Standards for School Mathematics*, NCTM, Reston, VA.

Pan, W. (2001) 'On the robust variance estimator in generalized estimating equations', *Biometrika*, Vol. 88, pp.901–906.

Patterson, B.H., Dayton, C.M. and Graubard, B.I. (2002) 'Latent class analysis of complex sample survey data: application to dietary data', *Journal of the American Statistical Association*, Vol. 97, No. 459, pp.721–741.

Quenouille, M.H. (1949) 'Problems in plane sampling', *The Annals of Mathematical Statistics*, Vol. 20, No. 3, pp.355–375.

Rabe-Hesketh, S. and Skrondal, A. (2006) 'Multilevel modeling of complex survey data', *Journal of the Royal Statistical Society (Series A)*, Vol. 169, No. 4, pp.805–827.

Ravand, H., Barati, H. and Widhiarso, W. (2013) 'Exploring diagnostic capacity of a high stakes reading comprehension test: a pedagogical demonstration', *Iranian Journal of Language Testing*, Vol. 3, No. 1, pp.1–27.

Robitzsch, A., Kiefer, T., George, A.C. and Uenlue, A. (2014) *CDM: Cognitive Diagnosis Modeling*, R Package Version 4.0. Available at: http://CRAN.R-project.org/package=CDM (accessed on 06-26-2017).

Tatsuoka, K. (1985) 'A probabilistic model for diagnosing misconceptions in the pattern classification approach', *Journal of Educational Statistics*, Vol. 10, No. 1, pp.55–73.

Templin, J.L. and Henson, R.A. (2006) 'Measurement of psychological disorders using cognitive diagnosis models', *Psychological Methods*, Vol. 11, pp.287–305.

von Davier, M. (2005) *A General Diagnostic Model Applied to Language Testing Data*, Research. Report RR-05-16, ETS, Princeton, NJ.

von Davier, M. (2014) 'The DINA model as a constrained general diagnostic model - two variants of a model equivalency', *British Journal of Mathematical and Statistical Psychology*, Vol. 67, pp.49–71.

Wang, M. and Long, Q. (2011) 'Modified robust variance estimator for generalized estimating equations with improved small-sample performance', *Statistics in Medicine*, Vol. 30, No. 11, pp.1298–1291.

Wedderburn, R.W.M. (1974) 'Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method', *Biometrika*, Vol. 61, No. 3, pp.439–447.

Wedel, M., Ter Hofstede, F. and Steenkamp, J.B.E.M. (1998) 'Mixture model analysis of complex samples', *Journal of Classification*, Vol. 15, No. 2, pp.225–244.

Wolter, K.M. (2007) *Introduction to Variance Estimation*, Springer Verlag, New York.

White, H. (1982) 'Maximum likelihood estimation of misspecified models', *Econometrica*, Vol. 50, pp.1–26.

Xu, X. and von Davier, M. (2008) *Fitting the Structured General Diagnostic Model to NAEP Data*, Research Report RR-08-27, Educational Testing Services, Princeton, NJ.

Yen, W.M. (1993) 'Scaling performance assessments: strategies for managing local item independence', *Journal of Educational Measurement*, Vol. 30, pp.187–213.

Appendix A: R code for Pan's method

```
dina.sand.strata.var.PAN<-function(inp,strata)
{
  require("MASS")
  if(class(inp)!="din") stop("Must provide an object of class
din")
  if(inp$rule!="DINA") stop("Input must be of result of fitted
DINA model")
  rule=inp$rule
  data<-inp$data
  if(nrow(data)!=length(strata)) stop("dimensions of strata and
data  must match")
  q<-inp$q.matrix
  I<-nrow(data)
  J<-ncol(data)
  K<-ncol(q)
```

```
  L<-2^K
  g<-inp$guess[,1]
  s<-inp$slip[,1]
  prob<-inp$posterior
  prob.all<-t(inp$attribute.patt)[1,,drop=F]
  w<-inp$weights
  attr.patt<-matrix(0,L,K)
  h1 <- 2
  if (K >= 2) {
    for (ll in 1:(K - 1)) {
      lk <- combn(1:K, ll)
      for (jj in 1:(ncol(lk))) {
        attr.patt[h1, lk[, jj]] <- 1
        h1 <- h1 + 1
      }
    }
  }

  attr.patt[L, ] <- rep(1, K)
  #create latresp or eta matrix
  comp <- (rowSums(q)) * (rule == "DINA")
  compL <- outer(comp, rep(1, L))
  attrpatt.qmatr <- t((attr.patt %*% t(q)))
  latresp <- apply(attr.patt, 1, FUN = function(attr.patt.ll) {
    attr.patt.ll <- outer(rep(1, J), attr.patt.ll)
    ind <- 1 * (rowSums(q * attr.patt.ll) >= comp)
    ind
    })

  #remove all zeroclasses form B, M and D
  allzeroclass<-c(1,getzeroclass(q))
  n.allzero<-length(allzeroclass)
  n.par<-L-n.allzero
  B<-M<-matrix(0,2*J,2*J)
  Bp<-Mp<-matrix(0,n.par,n.par)

  #strata Cov matrices
  unique.strata<-unique(strata)
  n.strata<-length(unique.strata)
  R.all<-cor(data,use="pairwise.complete.obs")
R.all[is.na(R.all)]<-0
#if no pairwise complete data, we assume them to be independent
  R.strata<-NULL
  CovY.all<-matrix(0,J,J)
  CovY.strata<-NULL
  for(i in 1: n.strata)
  {
    data.tmp<-data[which(strata==unique.strata[i]),]
```

```
    prob.tmp<-prob[which(strata==unique.strata[i]),]
    R.tmp<-cor(data.tmp,use="pairwise.complete.obs")
    R.tmp[is.na(R.tmp)]<-0
    R.strata[[i]]<-R.tmp
    strata.n<-nrow(data.tmp)
    CovY.strata[[i]]<-matrix(0,J,J)
    for(j in 1:strata.n)
    {
      pi<-prob.tmp[j,,drop=F]
      pihasskillforj<-as.vector(latresp%*%t(pi))
      Yi<-as.vector(as.numeric(data.tmp[j,]))
      mui<-as.vector((1-s)*pihasskillforj +g*(1-
pihasskillforj))
      invsqrtAi<-diag(1/sqrt(mui*(1-mui)))
      Yi[is.na(Yi)]<-0
      mui[is.na(mui)]<-0
      invsqrtAi[is.na(invsqrtAi)]<-0
      Yi<-matrix(Yi,J,1)
      mui<-matrix(mui,J,1)
      CovY.all<-CovY.all+invsqrtAi%*%(Yi-mui)%*%t(Yi-
mui)%*%invsqrtAi
      CovY.strata[[i]]<-CovY.strata[[i]]+invsqrtAi%*%(Yi-
mui)%*%t(Yi-mui)%*%invsqrtAi
    }
    CovY.strata[[i]]<-CovY.strata[[i]]/strata.n
  }
  CovY.all<-CovY.all/I

  for(i in 1:I)
  {
    Ri<-R.strata[[which(unique.strata==strata[i])]]
    CovYi<-CovY.strata[[which(unique.strata==strata[i])]]
    pi<-prob[i,,drop=F]
    pihasskillforj<-as.vector(latresp%*%t(pi))
    Dgi<-diag(1-pihasskillforj)
    Dsi<--diag(pihasskillforj)
    Dpi<-latresp*(1-s-g)
    Dpi<-Dpi[,-allzeroclass]
    Di<-cbind(Dgi,Dsi)
    Yi<-as.vector(as.numeric(data[i,]))
    mui<-as.vector((1-s)*pihasskillforj +g*(1-pihasskillforj))
    sqrtAi<-diag(sqrt(mui*(1-mui)))
    isNA<-which(is.na(Yi))
    if(length(isNA)>0)
    {
      Yi<-Yi[-isNA]
      mui<-mui[-isNA]
      Di<-Di[-isNA,,drop=F]
```

```
      Dpi<-Dpi[-isNA,,drop=F]
      Ri<-Ri[-isNA,-isNA,drop=F]
      sqrtAi<-sqrtAi[-isNA,-isNA,drop=F]
      CovYi<-CovYi[-isNA,-isNA,drop=F]
    }
    Vi<-sqrtAi%*%Ri%*%sqrtAi/w[i]
    Vi.inv<-ginv(Vi)
    CovYi<-sqrtAi%*%CovYi%*%sqrtAi
    Bi<-t(Di)%*%Vi.inv%*%Di
    Mi<-t(Di)%*%Vi.inv%*%CovYi%*%Vi.inv%*%Di
    B<-B+Bi
    M<-M+Mi
    Bpi<-t(Dpi)%*%Vi.inv%*%Dpi
    Mpi<-t(Dpi)%*%Vi.inv%*%CovYi%*%Vi.inv%*%Dpi
    Bp<-Bp+Bpi
    Mp<-Mp+Mpi
  }
  B.inv<-ginv(B)
  Bp.inv<-ginv(Bp)
  sand.cov<-B.inv%*%M%*%B.inv
  guess.sand.cov<-sand.cov[1:J,1:J]
  slip.sand.cov<-sand.cov[(J+1):(2*J),(J+1):(2*J)]
  attrpat.sand.cov<-Bp.inv%*%Mp%*%Bp.inv
  attr.patt.free<-attr.patt[-allzeroclass,]
  skillprob.sand.cov<-
t(attr.patt.free)%*%attrpat.sand.cov%*%attr.patt.free
  output<-inp
  output$guess.sand.cov=guess.sand.cov
  output$slip.sand.cov=slip.sand.cov
  output$attrpat.sand.cov=attrpat.sand.cov
  output$skillprob.sand.cov=skillprob.sand.cov
  output$strata=strata
  output$R.all=R.all
  output$R.strata=R.strata
  class(output)<-"din+strata+sandwich(DINA)"
  return(output)
}

getzeroclass<-function(q)
{
  J<-nrow(q)
  K<-ncol(q)
  L<-2^K
  attr.patt<-matrix(0,L,K)
  h1 <- 2
  if (K >= 2) {
    for (ll in 1:(K - 1)) {
      lk <- combn(1:K, ll)
```

```
      for (jj in 1:(ncol(lk))) {
        attr.patt[h1, lk[, jj]] <- 1
        h1 <- h1 + 1
      }
    }
  }
  attr.patt[L, ] <- rep(1, K)
  #create latresp or eta matrix
  comp <- (rowSums(q))
  compL <- outer(comp, rep(1, L))
  attrpatt.qmatr <- t((attr.patt %*% t(q)))
  latresp <- apply(attr.patt, 1, FUN = function(attr.patt.ll) {
    attr.patt.ll <- outer(rep(1, J), attr.patt.ll)
    ind <- 1 * (rowSums(q * attr.patt.ll) >= comp)
    ind
  })
  itempossible<-colSums(latresp)
  zeroclass.t<-which(itempossible==0)
  zeroclass<-NULL
  if(length(zeroclass.t)>1) zeroclass<-zeroclass.t[-1]
  return((zeroclass))
}

result<-dina.sand.strata.var.PAN(din(data=data,q.matrix
=q,zeroprob.skillclasses=getzeroclass(q),weights=data1[,2],rule
="DINA",conv.crit=.00001,progress=F),strata=strata)
```

Appendix B: Q-matrix for TIMSS 2011 grade 8 math data

Domain		*Number concepts*		*Algebra concepts*		*Geometry concepts*			*Data and probability*	
Block	*Item*	*Whole numbers and integers*	*Fractions, decimals and proportions*	*Patterns*	*Expressions, equations and functions*	*Lines, angles and shapes*	*Measurement*	*Location and movement*	*Data organisation, representation and interpretations*	*Probability*
M01	1	1	0	0	0	0	0	0	0	0
M01	2	0	0	1	0	0	0	0	1	0
M01	3	1	0	1	0	0	0	0	1	0
M01	4a	1	0	1	0	0	0	0	1	0
M01	4b	1	0	1	0	0	0	0	1	0
M01	4c	1	0	1	0	0	0	0	1	0
M01	5	0	0	1	1	0	0	0	0	0
M01	6	0	0	0	0	1	0	0	0	0
M01	7	1	0	0	0	0	0	0	0	0

Appendix B: Q-matrix for TIMSS 2011 grade 8 math data (continued)

Domain		*Number concepts*		*Algebra concepts*		*Geometry concepts*			*Data and probability*	
Block	*Item*	*Whole numbers and integers*	*Fractions, decimals and proportions*	*Patterns*	*Expressions, equations and functions*	*Lines, angles and shapes*	*Measurement*	*Location and movement*	*Data organisation, representation and interpretations*	*Probability*
M01	8	0	1	0	0	0	0	0	1	0
M01	9	0	0	0	1	0	0	0	0	0
M02	1	0	1	0	0	0	0	0	0	0
M02	2	0	1	0	0	0	0	0	0	0
M02	3	1	0	0	0	0	0	0	0	0
M02	4	0	1	0	0	0	0	0	0	0
M02	5	0	1	0	0	0	0	0	0	0
M02	6	0	0	0	1	0	1	0	0	0
M02	7	1	0	0	1	0	0	0	0	0
M02	8	1	0	0	1	0	0	0	0	0
M02	9	0	0	0	0	0	1	0	0	0
M02	10	0	0	0	0	1	0	0	0	0
M02	11	1	0	0	0	0	1	0	0	0
M02	12	0	0	0	0	0	1	1	0	0
M02	13	0	0	0	0	0	0	0	0	1
M02	14a	0	0	0	0	0	0	0	1	0
M02	14b	0	0	0	0	0	0	0	1	0
M03	1	0	1	0	0	0	0	0	0	0
M03	2	0	1	0	0	0	0	0	0	0
M03	3	1	0	1	0	0	0	0	0	0
M03	4	0	1	0	0	0	0	0	0	0
M03	5	0	0	0	1	0	0	0	0	0
M03	6	1	0	0	1	0	0	0	0	0
M03	7	1	0	0	1	0	0	0	0	0
M03	8	1	0	0	1	0	0	0	0	0
M03	9	1	0	1	0	0	0	0	0	0
M03	10	1	0	0	1	0	0	0	0	0
M03	11	0	0	0	0	1	0	0	0	0
M03	12	1	0	0	0	0	1	0	0	0
M03	13	0	0	0	0	0	0	1	0	0
M03	14	0	0	0	0	0	0	0	1	0
M03	15	0	0	0	0	0	0	0	0	1
M03	16	0	1	0	0	0	0	0	0	0
M03	17	0	1	0	0	0	0	0	1	0

Appendix B: Q-matrix for TIMSS 2011 grade 8 math data (continued)

Domain		*Number concepts*		*Algebra concepts*		*Geometry concepts*			*Data and probability*	
Block	*Item*	*Whole numbers and integers*	*Fractions, decimals and proportions*	*Patterns*	*Expressions, equations and functions*	*Lines, angles and shapes*	*Measurement*	*Location and movement*	*Data organisation, representation and interpretations*	*Probability*
M05	1	0	1	0	0	0	0	0	0	0
M05	2	0	1	0	0	0	0	0	0	0
M05	3	0	1	0	0	0	0	0	0	0
M05	4	0	0	0	1	0	0	0	0	0
M05	5	0	0	0	1	0	0	0	0	0
M05	6	1	0	0	1	0	0	0	0	0
M05	7	0	1	0	0	1	0	0	0	0
M05	8	1	0	0	0	0	1	0	0	0
M05	9	0	0	0	0	0	0	1	0	0
M05	10	0	0	0	0	1	0	0	0	0
M05	11	0	0	0	0	0	0	1	0	0
M05	12	0	0	0	0	0	0	1	0	0
M05	13	0	1	0	0	0	0	0	1	0
M05	14	0	0	0	0	0	0	0	0	1
M06	1	0	1	0	0	0	0	0	0	0
M06	2	0	1	0	0	0	0	0	0	0
M06	3	1	0	0	0	0	0	0	1	0
M06	4	1	0	0	0	0	0	0	0	0
M06	5a	0	0	1	0	0	0	0	0	0
M06	5b	0	0	1	0	0	0	0	0	0
M06	5c	0	0	1	0	0	0	0	0	0
M06	6	1	0	0	1	0	0	0	0	0
M06	7	0	0	0	1	0	0	0	0	0
M06	8	0	0	0	1	0	1	0	0	0
M06	9	0	0	0	0	1	0	0	0	0
M06	10A	0	0	0	0	0	0	1	1	0
M06	10B	0	0	0	0	1	0	1	1	0
M06	11	0	0	0	0	0	0	0	0	1
M06	12a	0	0	0	0	0	0	0	1	0
M06	12b	0	0	0	0	0	0	0	1	0
M06	12c	0	0	0	0	0	0	0	1	0
M07	1	1	0	1	0	0	0	0	1	0
M07	2	0	1	0	0	0	0	0	0	0
M07	3	0	1	0	1	0	0	0	0	0

Appendix B: Q-matrix for TIMSS 2011 grade 8 math data (continued)

sDomain		*Number concepts*		*Algebra concepts*		*Geometry concepts*			*Data and probability*	
Block	*Item*	*Whole numbers and integers*	*Fractions, decimals and proportions*	*Patterns*	*Expressions, equations and functions*	*Lines, angles and shapes*	*Measurement*	*Location and movement*	*Data organisation, representation and interpretations*	*Probability*
M07	4	0	0	0	1	0	0	0	0	0
M07	5	0	0	0	1	0	0	0	0	0
M07	6	0	0	0	~~0~~	1	0	1	0	0
M07	7	1	0	0	0	0	1	0	0	0
M07	8	0	0	0	0	1	0	1	0	0
M07	9	1	0	0	1	0	0	0	0	0
M07	10	0	0	0	0	1	0	0	0	0
M07	11	0	0	0	0	0	0	0	0	1
M07	12	0	0	0	1	0	0	0	~~0~~	0
M07	13a	0	0	0	0	0	0	0	1	0
M07	13b	0	0	0	0	0	0	0	1	0
M07	13c	0	0	0	0	0	0	0	1	0