Multivariate Behavioral Research, 49:54–66, 2014

ISSN: 0027-3171 print / 1532-7906 online
DOI: 10.1080/00273171.2013.843441

The Joint Multivariate Modeling of Multiple Mixed Response Sources: Relating Student Performances with Feedback Behavior

Jean-Paul Fox
University of Twente, The Netherlands

Rinke Klein Entink
TNO Zeist, The Netherlands

Caroline Timmers
Saxion University of Applied Sciences, The Netherlands

The present study concerns a Dutch computer-based assessment, which includes an assessment process about information literacy and a feedback process for students. The assessment is concerned with the measurement of skills in information literacy and the feedback process with item-based support to improve student learning. To analyze students' feedback behavior (i.e. feedback use and attention time), test performance, and speed of working, a multivariate hierarchical latent variable model is proposed. The model can handle multivariate mixed responses from multiple sources related to different processes and comprehends multiple measurement components for responses and response times. A flexible within-subject latent variable structure is defined to explore multiple individual latent characteristics related to students' test performance and feedback behavior. Main results of the computer-based assessment showed that feedback-information pages were less visited by well-performing students when they relate to easy items. Students' attention paid to feedback was positively related to working speed but not to the propensity to use feedback.

Different advantages of computer-based assessment (e.g., improved reliability, adaptive testing, use of multimedia technology) have been exploited the last 2 decades (e.g., van der Linden & Glas, 2010). Recently, the possibility of providing students instant feedback has been further explored. Different computer-based assessment studies have investigated the effects of feedback on student learning (e.g., Hattie & Timperly, 2007; van der Kleij, Eggen, Timmers, & Veldkamp, 2012). Individual homogeneity in the use of feedback is generally assumed, but it can be expected that students differ in their feedback use. For example, Timmers & Veldkamp (2011) showed that the attention paid to feedback differs over students, where student characteristics such as study motivation and a positive attitude may influence the time spent reading feedback (Aleven, Stahl, Schworm, Fischer, & Wallace, 2003; van der Kleij, Eggen, & Timmers, 2011). Heterogeneity in feedback use should be taken into account to obtain accurate effects of feedback on student learning. Furthermore, differential effects of feedback on student learning can be studied by conditioning on feedback behavior.

In the present computer-based assessment study, 1st-year bachelor students of a university of applied sciences could consult elaborate feedback after completing an information literacy test. After they finished the test, a knowledge-of-result page was automatically generated, which showed the correctness of each response and links to pop-up pages with additional feedback per item. Each feedback page showed the correct response and an explanation of various concepts used in the item and answering categories. Besides the responses and response times to the test items, the system stored whether or not a pop-up page was opened. Furthermore, the time from opening to closing a pop-up page was also recorded.

Correspondence concerning this article should be addressed to Jean-Paul Fox, Faculty of Behavioral Sciences, University of Twente, The Netherlands. E-mail: j.p.fox@utwente.nl

The assessment is carried out to measure students' proficiency level and to identify where specific deficiencies exist. When missing skills are identified, relevant feedback can be provided to help students build necessary skills. This is arranged by the process that generates feedback using an information retrieval system (e.g., Cool & Belkin, 2011), where students can obtain relevant information by clicking through the system.

The information retrieval system supports information-seeking behavior of the student, where the object is to satisfy the information need of the student. However, it contains all stored feedback information for all students, and optionally with a search engine such as Google, an enormous amount of feedback can be provided. This makes it difficult for students to obtain easily relevant feedback information. Furthermore, although assessment results may show a lack of knowledge, the explicit information needs can still be vague. Interpretation problems with the obtained assessment results can also complicate the search for relevant feedback.

To better understand students' information-seeking behavior, implicit feedback measures such as click-through rates and pop-up opening times are observed. The idea is that by learning about preferences of students using an in-depth analysis of students' seeking behavior, the efficiency and accuracy of the system's feedback information can be further improved and used to individuate system responses. Therefore, a latent variable modeling approach is carried out to analyze students' seeking behavior using implicit feedback measures: (a) feedback use given click-through data and (b) attention given the pop-up opening times. It is typically assumed that students open feedback pages that are relevant and time spent viewing a page reveals the time to process the feedback information and the selective attention of the students.

Although seemingly easy to monitor, implicit feedback observations are difficult to interpret and potentially ambiguous. Therefore, latent variables are used to model unobserved within- and between-student heterogeneity in feedback use and attention time, to estimate variance components, and to account for measurement error. Furthermore, latent variables feedback use and attention are related to the assessment measurements ability and working speed to explore and validate the usefulness of the feedback behavior measurements in providing relevant feedback information.

The proposed Bayesian joint modeling framework captures the cross-classified structure of the data, where two different levels can be defined because responses are nested in persons and items. This approach is based on recent work in joint modeling of item responses and response times (e.g., Fox, 2010; Klein Entink, Fox, & van der Linden, 2009; van der Linden, 2007; C. Wang, Chang, & Douglas, 2013; T. Wang & Hanson, 2005). In the present approach, four different measurement models are considered for four different latent variables using different observed response types. A joint modeling approach is pursued, where item observations are clustered within latent variables. Subsequently, the four latent variables are assumed to be multivariate normally distributed and allowed to correlate within the student.

The item characteristics of the measurement models are also allowed to correlate. For example, item thresholds related to feedback use may correlate with thresholds related to student achievement and also with time intensities related to speed of working. As a specific case, item difficulties can be negatively correlated with thresholds for feedback use, which indicates that difficult items induce positive feedback behavior.

Note that the joint model can handle multiple different types of responses, which are collected for each student from both processes. The assessment data consist of discrete test scores and continuous response times, which are indicators of ability and speed of working, respectively. The feedback behavior data represent click-through data and attention times and an indicator variable is defined indicating the use or disuse of feedback. This supports a manifest mixture modeling approach such that the feedback behavior data are conditionally modeled on feedback use.

The article is organized as follows: After the description of the multivariate measurement model, a Markov chain Monte Carlo (MCMC) algorithm is proposed, which represents the sampling steps related to the different measurement components. A small simulation study illustrates the parameter recovery of the algorithm. Subsequently, the model is applied to the Dutch feedback data. Finally, the last section provides a discussion and outlines directions for further research.

STUDENT PERFORMANCE AND FEEDBACK BEHAVIOR

The computer-adaptive process consists of two stages. In stage one, students are assessed and test results are used to construct a knowledge-of-results page. This page contains links to additional feedback information such that students can search and retrieve elaborate feedback information from the system. In this second stage, the system, which provides the feedback, can be informed about students' feedback-seeking behavior.

The click-through data consisting of zeros and ones, items for which feedback is not and is consulted, respectively, are indicators of the latent variable feedback use. This latent variable represents the propensity of consulting feedback. For each item, the time that an elaborated feedback page is opened is considered the reading or attention time and is used to measure the latent variable attention. The attention times relate to the ability to attend and to process specific information and to discard irrelevant information. This corresponds to the psychological concept of executive attention, which governs selected actions and order of the actions. Students with high attention levels can more easily focus on specific shortcomings and process the feedback

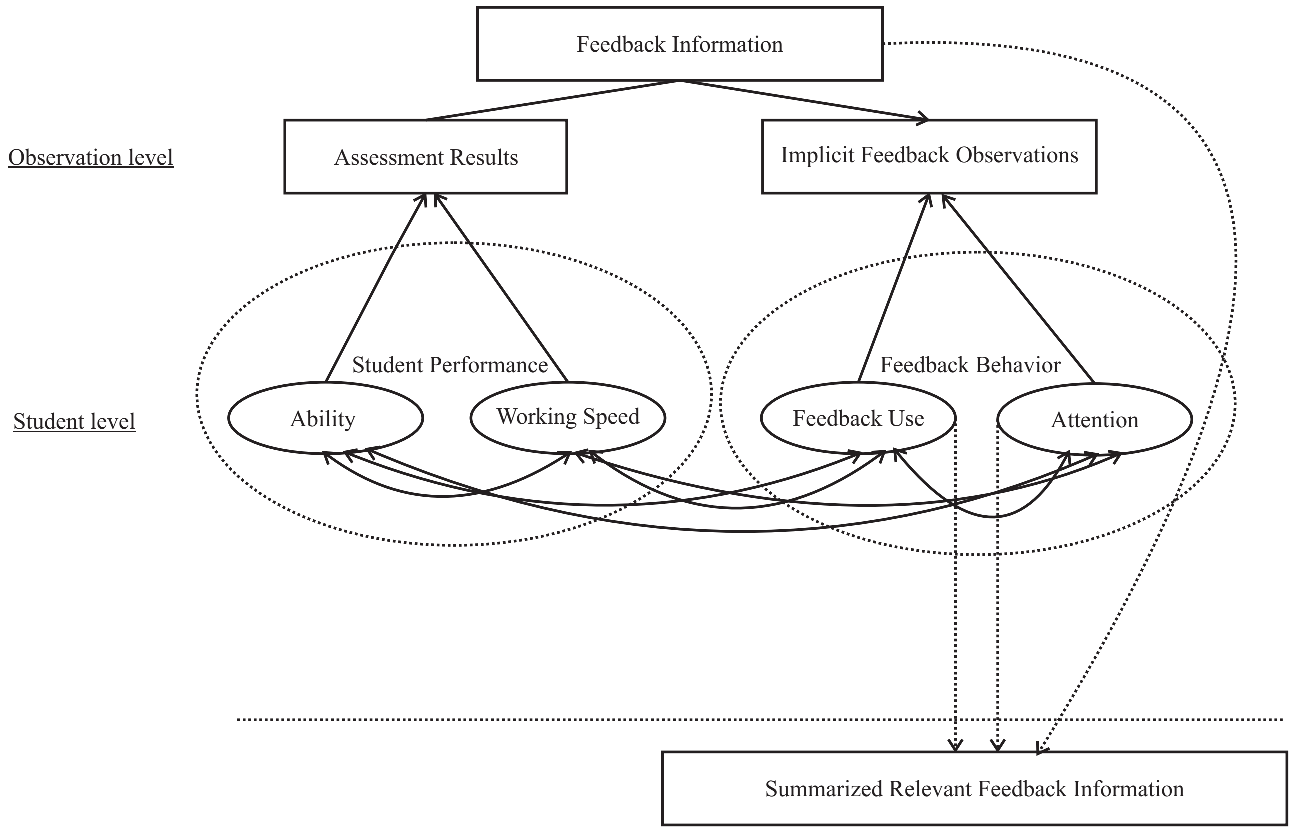

FIGURE 1 Computer-based assessment with automatically generated feedback: a schematic overview of the joint hierarchical modeling framework (solid lines).

information, and they show lower attention times than those with low attention levels.

The processes are displayed in Figure 1. The left-hand side shows the latent variable framework concerning the assessment, where ability and working speed are underlying the test results. The test results are used to generate feedback information, where links are given to more elaborate feedback. Students are given the opportunity to search the information retrieval system. The click-through data and attention time data are stored as implicit feedback observations, which represent students' feedback behavior. The attention times are conditionally modeled on positive feedback use observations.

The latent variable framework is meant for an exploratory analysis, where interest is focused on between-student construct variability and relationships between constructs. First, it is investigated to what extent students seek information when they are invited to use the information retrieval system. Subsequently, interest is focused on the difference in intensity to search feedback between low- and high-performing students. Furthermore, attention levels of high-performing students and low-performing students are compared. Second, the variability in student feedback behavior (i.e., feedback use and attention) is explored to investigate differences in the use of automatically generated feedback and whether student ability and working speed influence feedback behavior. The developed information retrieval system is probably not the optimal support tool for all students. Therefore, it is investigated who are actively searching for relevant feedback. It might also be possible to identify atypical individuals with respect to their ability, speed, and/or feedback behavior. Third, interest is focused on the differential use of feedback given students' overall test performance and working speed. When feedback behavior correlates with student performance, it can be expected that the feedback behavior can be used to define more efficiently and accurately student-specific feedback, which is represented via dotted lines in the schematic overview in Figure 1. Then, in a more interactive way, automatically generated feedback can be improved by taking into account the feedback behavior of students.

MULTIVARIATE MULTILEVEL LATENT VARIABLE MODELING

Each item, index $k = 1, \ldots, K$, refers to a test item k but also to its feedback page. A total of four observations are associated with item k: the response observation, feedback use, time to solve the item, and the feedback attention time. The four blocks of multivariate item response data are assumed to be indicators of four different latent variables all defined at the person level. Student ability and speed of working are measured using the assessment data consisting of discrete binary responses and continuous response times.

In the proposed multivariate latent variable approach, it is assumed that within-subject observations are conditionally independently distributed given the latent variable. Therefore, two item response theory models describe the distributions of the binary responses related to the achievement test and feedback items. Furthermore, two item response time models define the distribution of the log response times associated with the responses to the achievement test and feedback items.

At a higher level, two covariance structures are defined to model dependencies between the Level 1 model parameters. Within a student, the latent variable ability, speed of working, feedback use, and attention are assumed to be correlated. The covariance structure is assumed to be common across students. A speed-accuracy trade-off is assumed such that accuracy is influenced when changing the speed of working (e.g., van der Linden, 2009). The constructs feedback use and attention do not follow a trade-off, where attention times are also conditionally modeled on feedback use. The task of retrieving relevant information from a system requires different skills, and different strategies are possible to succeed. For example, retrieving relevant information can be accomplished by opening many feedback pages (high level of feedback use) and detailed reading of the material (high level of attention). It is also possible that an organized student effectively opens relevant feedback pages (low level of feedback use) and processes the relevant information quickly (low level of attention).

A second covariance structure is defined at the level of items. The item characteristics of the test items and the feedback items are assumed to be correlated. This means that the item characteristics related to the test and speed performances can correlate. An item characteristic analysis can be used to explore relationships between item difficulties and feedback thresholds. For example, characteristics of popular feedback items can be related to specific test item characteristics. Furthermore, the feedback-time intensities, representing the average attention time, can be related to the item difficulties and the feedback thresholds. The characteristics of items, from response and response time information, that stimulate feedback use are of interest. These items can be used to construct a test that supports the use of feedback and/or feedback learning.

Recently, survival models have been proposed to measure latent traits using response times (e.g., Ranger & Ortner, 2011). C. Wang et al. (2013) proposed the general linear transformation model, which includes Cox proportional hazards model and other parametric models, among others. Bianconcini & Cagnone (2012) proposed a multivariate latent growth model to analyze mixed multivariate individual responses from different sources but all related to student performance. This approach is based on the generalized linear latent variable model of Moustaki & Knott (2000) for mixed responses in the exponential family. From a Bayesian modeling and joint estimation perspective, it turns out to be convenient to use the log-normal model for response times. In the response time modeling literature, the log-normal model for response times on test items is widely accepted, which allows for a speed-accuracy trade-off and item-specific time-intensive parameters (van der Linden, 2006).

Measurement Models Related to Ability and Speed of Working

Let the person parameters of a student indexed $i = 1, \ldots, N$, for ability and speed be denoted as θ_i^a and ζ_i^a, respectively, where the superscript a refers to the achievement test. The two other person parameters, propensity to use feedback and feedback attention, are denoted as θ_i^f and ζ_i^f, respectively, where the superscript f refers to the feedback process.

Item response models are defined to measure the latent variables ability and feedback use. For the achievement test items, the probability of student i answering item k correctly ($Y_{ik}^a = 1$) is assumed to follow the three-parameter item response model:

$$\begin{aligned} P\left(Y_{ik}^a = 1 \mid \theta_i^a, a_k, b_k, c_k\right) \\ = c_k + (1 - c_k)\Phi\big(a_k\big(\theta_i^a - b_k\big)\big), \end{aligned} \tag{1}$$

where $\Phi(.)$ denotes the normal cumulative distribution function and a_k, b_k, and c_k the discrimination, difficulty, and guessing parameters of item k, respectively.

For the feedback component, the probability that student i opens item k's feedback page ($Y_{ik}^f = 1$) is assumed to follow the two-parameter item response model:

$$P\big(Y_{ik}^f = 1 \mid \theta_i^f, \alpha_k, \beta_k\big) = \Phi\big(\alpha_k\big(\theta_i^f - \beta_k\big)\big), \tag{2}$$

where α_k and β_k are the discrimination and threshold parameters of feedback item k, respectively. The two-parameter model is used because guessing does not play a role in feedback use.

The response times are restricted to be positive. Therefore, the log-response time of student i on item k of the achievement test, denoted as T_{ik}^a, is assumed to be normally distributed:

$$T_{ik}^a \sim \mathcal{N}\big(h_k - g_k\zeta_i^a, \sigma_k^{2(a)}\big), \tag{3}$$

where h_k and g_k are the time intensity and the time discrimination parameters. The mean log-response time is a function of the speed of working, which is represented by the latent variable ζ_i^a. An increase in working speed is represented by an increase in the latent variable, which leads to a lower response time. When the time intensity is increased the item is expected to consume more time. The time discrimination parameter makes it possible that, without violating the stationary speed assumption, a faster working student may need less time for an item than would be expected on the basis of his or her speed of working. This effect corresponds to the discriminative character of items between students with different abilities.

Mixture Measurement Models Related to Feedback and Attention

Observed attention times are fixed to zero for those items for which feedback pages were not consulted. Therefore, attention times are conditionally modeled on observed feedback use using a mixture model, which is represented by

$$p\left(T_{ik} \mid \zeta_i^f, \lambda_k, \phi_k, Y_{ik}^f\right) = P\left(Y_{ik}^f = 0\right) P(T_{ik} = 0) + P\left(Y_{ik}^f = 1\right) p\left(T_{ik} \mid \lambda_k, \phi_k, \zeta_i^f\right).$$

Conditional on using feedback, the attention times are modeled conditional on the propensity to use feedback, and the time dealing with feedback is naturally restricted to be positive. In the same way, student i's log attention time of item k's feedback, denoted as T_{ik}^f, is assumed to be normally distributed:

$$T_{ik}^f \mid Y_{ik}^f = 1 \sim \mathcal{N}\left(\lambda_k - \phi_k \zeta_i^f, \sigma_k^{2(f)}\right), \tag{4}$$

where λ_k and ϕ_k are the time intensity and the time discrimination parameters of the feedback item k. The latent variable, ζ_i^f, represents the student's level of attention. The time intensity parameter represents the average logarithm of attention time when the student's attention level equals zero. The time discrimination parameter captures the variability in attention times over items given a constant level of attention.

The mean structures in Equation (3) and Equation (4) are defined in such a way that a high level of speed of working leads to low response times and a high attention level leads to low attention times.

The data augmentation scheme of Albert (1992) and Albert & Chib (1993) was used to augment the binary response observations with latent continuous responses Z_{ik}^a and Z_{ik}^f. The augmented data are independently and normally distributed with variance one and truncated to be positive when Y_{ik}^a and Y_{ik}^f are greater than zero, respectively. This supports a straightforward implementation of an MCMC algorithm, where mean structure parameters can be easily sampled from normal distributions (see Appendix). Given the augmented data, the vector of normally distributed outcomes of student i to test and feedback item k can be represented as

$$\begin{bmatrix} Z_{ik}^a \\ Z_{ik}^f \\ T_{ik}^a \\ T_{ik}^f \mid Y_{ik}^f = 1 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} a_k(\theta_i^a - b_k) \\ \alpha_k(\theta_i^f - \beta_k) \\ g_k - h_k\zeta_i^a \\ \lambda_k - \phi_k \zeta_i^f \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \sigma_k^{2(a)} & 0 \\ 0 & 0 & 0 & \sigma_k^{2(f)} \end{bmatrix}\right), \tag{5}$$

when the student used the feedback of item k and did not guess the response to test item k. The within-student outcomes are assumed to be independently distributed given the person parameters. Students might also guess. Therefore, a latent response variable S_{ik} can be introduced, which is equal to one when the student i knows the response and zero when the item is guessed correctly. In that case, the distribution of Z_{ik}^a will depend on the value of S_{ik} (e.g., Béguin & Glas, 2001; Fox, 2010).

To account for respondents who will not consult any of the feedback pages, a mixture model is defined for the distribution of feedback-use data. Let the random variable G_i define whether participant i belongs to the group who consult feedback by opening a least one feedback page ($G_i = 1$), referred to as the feedback group, or will not open any of the feedback pages ($G_i = 0$), referred to as the nonfeedback group. The feedback-use data of respondent i are distributed as

$$p\left(\mathbf{y}_i^f \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta_i^f\right) = P(G_i = 1)\, p\left(\mathbf{y}_i^f \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta_i^f\right) + P(G_i = 0)\, p\left(\mathbf{y}_i^f = \mathbf{0}\right). \tag{6}$$

To avoid the complex modeling task, where the distributional parameters of continuous latent variables depend on latent class parameters, a manifest mixture approach is pursued. When at least one feedback page is consulted, student i is classified to the feedback group with probability one. Otherwise, student i is classified with probability one to the nonfeedback group. As a result, the measurement parameters related to feedback are based on observations from participants using feedback, where the measurement parameters related to ability are based on all participant information.

Multivariate Model for the Person Parameters

At the level of students, the four person parameters $(\theta_i^a, \theta_i^f, \zeta_i^a, \zeta_i^f)$ are assumed to be multivariate normally distributed. A common mean $\boldsymbol{\mu}_P$ and covariance structure is assumed over persons $\boldsymbol{\Sigma}_P$. The components of the mean vector represent the average level of achievement, feedback use, working speed, and attention in the population.

The diagonal terms of the covariance matrix $\boldsymbol{\Sigma}_P$ are denoted as $\sigma^2_{\theta^a}, \sigma^2_{\theta^f}, \sigma^2_{\zeta^a}$, and $\sigma^2_{\zeta^f}$, respectively. Each component describes the variation of a latent student characteristic in the population. The nondiagonal terms represent the covariances between the latent student characteristics. At Level 2, the multivariate latent variable model for the person parameters is represented by

$$\begin{bmatrix} (\theta_i^a, \zeta_i^a)^t \\ (\theta_i^f, \zeta_i^f)^t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_P^a \\ \boldsymbol{\mu}_P^f \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{P_{11}} & \boldsymbol{\Sigma}_{P_{12}} \\ \boldsymbol{\Sigma}_{P_{21}} & \boldsymbol{\Sigma}_{P_{22}} \end{bmatrix}\right), \tag{7}$$

where $\boldsymbol{\mu}_P^a$ and $\boldsymbol{\mu}_P^f$ denote the latent population means of ability and speed of working and of feedback use and feedback attention time, respectively. The covariance matrix $\boldsymbol{\Sigma}_P$ is partitioned to define the covariance matrix of ability and speed, $\boldsymbol{\Sigma}_{P_{11}}$; the covariance matrix of feedback use and reading,

$\boldsymbol{\Sigma}_{P_{22}}$; and the covariance matrix between these two blocks, $\boldsymbol{\Sigma}_{P_{12}}$.

Multivariate Model for the Item Parameters

A multivariate normal distribution for the item characteristics is complicated due to the fact that the guessing parameters are restricted to the interval [0, 1]. Therefore, define $\tilde{c}_k = \Phi^{-1}(c_k)$ such that $\tilde{c}_k$ is the normal deviate with $\Phi^{-1}(.)$ the inverse of the cumulative normal distribution function. The monotone transformation function defines a real-valued $\tilde{c}_k$ in a one-to-one correspondence with c_k (Fox, 2010).

A multivariate normal prior distribution can be specified for the set of item parameters of test and feedback item k; that is,

$$\begin{bmatrix} a_k \\ b_k \\ \tilde{c}_k \\ \alpha_k \\ \beta_k \\ g_k \\ h_k \\ \phi_k \\ \lambda_k \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_a \\ \mu_b \\ \mu_{\tilde{c}} \\ \mu_\alpha \\ \mu_\beta \\ \mu_g \\ \mu_h \\ \mu_\phi \\ \mu_\lambda \end{bmatrix}, \right.$$

$$\left. \begin{bmatrix} \begin{vmatrix} \tau_a^2 & \sigma_{ab} & \sigma_{a\tilde{c}} \\ \sigma_{ab} & \tau_b^2 & \sigma_{b\tilde{c}} \\ \sigma_{a\tilde{c}} & \sigma_{b\tilde{c}} & \tau_{\tilde{c}}^2 \end{vmatrix} & \cdots & \cdots & \cdots \\ \cdots & \begin{vmatrix} \tau_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \tau_\beta^2 \end{vmatrix} & \cdots & \cdots \\ \cdots & \cdots & \begin{vmatrix} \tau_g^2 & \sigma_{gh} \\ \sigma_{gh} & \tau_h^2 \end{vmatrix} & \cdots \\ \cdots & \cdots & \cdots & \begin{vmatrix} \tau_\phi^2 & \sigma_{\phi\lambda} \\ \sigma_{\phi\lambda} & \tau_\lambda^2 \end{vmatrix} \end{bmatrix} \right).$$

The four blocks represent the covariance structure of the item characteristics of the four measurement models. The first within-item block structure specifies the covariance between the item characteristics related to the test items for measuring achievement. The other blocks correspond to the measurement of feedback use, speed of working, and level of attention, respectively. The elements within a block are allowed to correlate but also between blocks. Although the off-diagonal blocks were not specified in the aforementioned equation, a full covariance matrix is assumed that allows for within-block and between-block correlations.

The multivariate model for the parameters of item k can be represented as the vector $\boldsymbol{\xi}_k^a = (a_k, b_k, \tilde{c}_k, \alpha_k, \beta_k)^t$ related to the achievement test and $\boldsymbol{\xi}_k^f = (g_k, h_k, \phi_k, \lambda_k)^t$ related to the feedback test. Subsequently, the vectors are multivariate normally distributed,

$$\begin{bmatrix} \boldsymbol{\xi}_k^a \\ \boldsymbol{\xi}_k^f \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_I^a \\ \boldsymbol{\mu}_I^f \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{I_{11}} & \boldsymbol{\Sigma}_{I_{12}} \\ \boldsymbol{\Sigma}_{I_{21}} & \boldsymbol{\Sigma}_{I_{22}} \end{bmatrix} \right), \tag{8}$$

where $\boldsymbol{\mu}_I^a$ and $\boldsymbol{\mu}_I^f$ denote the item parameter population means of the component ability and the component feedback, respectively. The covariance matrix of all item parameters, $\boldsymbol{\Sigma}_I$, is partitioned to define the covariance matrix of the item parameters for the ability part, $\boldsymbol{\Sigma}_{I_{11}}$; for the feedback part, $\boldsymbol{\Sigma}_{I_{22}}$; and the covariance matrix between the two blocks, $\boldsymbol{\Sigma}_{I_{12}}$.

The covariance structure can provide meaningful insights. It may be expected that easy items require less time to be solved than more difficult items. Then, the time intensity parameter correlates positively with the item difficulty parameter of the test item. The relationship between the item characteristics of feedback use and times may show a positive relation between time intensity and difficulty. But cross-correlations with the characteristics of the test items are also possible. The guessing parameter of the response model for the test items has no analogous parameter in the other measurement models (because there is no guessing aspect for response time or feedback use).

Model Identification and Parameter Estimation

The joint model representing the measurement of four latent variables together with a higher level latent covariance structure can be characterized as a multivariate Response Time Item Response Theory (RTIRT; Klein Entink et al., 2009; van der Linden, 2007). Two blocks (assessment and feedback component) of responses and response times are modeled, which leads to the multivariate extension of the RTIRT model.

The multivariate latent variable model is identified by fixing the latent multiple scales. In the present approach, the covariance matrix of the latent person parameters is restricted, where the variance parameter of ability and of feedback use is set to one. The variance restrictions identify the variance of the ability and feedback-use scale. It is also possible to set each product of discriminations equal to one. For each scale, the sum of the item difficulties are set to zero to identify the mean of each scale.

The mean population level of speed of working and of reading is set to zero. Furthermore, for each response time model, the product of time discrimination parameters is set to one to identify the variance of the scale. The total set of restrictions identifies the metric of the four different scales. The four latent scales are linked using a joint estimation procedure.

MCMC is used to estimate the model parameters. A summarized description is given in the Appendix. The scheme contains the full conditional posterior distributions from which samples are drawn consecutively. References are given to well-known sampling steps.

The MCMC algorithm is programmed in R and is freely available via the authors' website. In the program, the inverse-Wishart hyperprior for the covariance parameters ($\boldsymbol{\Sigma}_P$ and $\boldsymbol{\Sigma}_I$) has a diagonal scale parameter and minimum degrees of freedom. The normal hyperpriors for the mean parameters ($\boldsymbol{\mu}_P$ and $\boldsymbol{\mu}_I$) have zero means. The hyperpriors for the discriminations and time discriminations have a mean of one. Large variances are specified, which represent vague prior information.

PARAMETER RECOVERY STUDY

A simulation study was performed to investigate parameter recovery and sample size sensitivity. Different simulation conditions were considered, where special focus was thereby on the accuracy of the covariance between person parameter estimates.

Item parameters were simulated from independent normal distributions, where all discrimination and time discrimination parameters were drawn from an $N(1, .10)$ distribution and all difficulty and intensity parameters from a standard normal distribution. The guessing parameters for the measurement model for ability were all set to zero in this simulation.

Each set of person parameters, $(\theta^a, \zeta^a, \theta^f, \zeta^f)$, were generated from a multivariate normal distribution with $\boldsymbol{\mu}_P = \mathbf{0}$ and correlation matrix

$$\boldsymbol{\Sigma}_P = \begin{bmatrix} 1 & 0 & \rho_1 & \rho_2 \\ 0 & 1 & 0 & 0 \\ \rho_1 & 0 & 1 & \rho_3 \\ \rho_2 & 0 & \rho_3 & 1 \end{bmatrix}. \quad (9)$$

Zero to moderate correlations were specified and data sets of moderate size were generated and correspond to the setting of the real data study. Despite the moderate data samples, it is shown that still reasonably accurate parameter estimates were obtained. Therefore, different correlation structures were simulated using the restriction $\rho_1 = -\rho_2 = -2\rho_3$ with $\rho_1 = 0$ and $\rho_1 = -.50$. Subsequently, generated item and person parameters were used to simulate responses, response times, feedback use, and feedback times. This was done for $K = 10$ and $K = 20$ items and for $N = 300$ and $N = 500$ persons. For each combination, 50 replications were generated.

The MCMC algorithm was run for $3,000$ iterations and the first 500 draws were discarded as the burn-in period. In Table 1, the estimated correlations (ρ_1, ρ_2, ρ_3) and standard deviations are presented. As can be seen from Table 1, the simulated correlations are accurately estimated under different conditions. Also, the estimation of item and person parameters was investigated and showed a good correspondence between simulated and estimated parameters (not shown for reasons of space requirements).

TABLE 1
For Different Design Conditions, Estimated Correlations Between Person Parameters and Standard Deviations Averaged Over 50 Replications

		ρ_1			ρ_2			ρ_3		
K	N	True	EAP	SD	True	EAP	SD	True	EAP	SD
10	300	.00	.00	.04	.00	.00	.04	.00	.01	.04
	500		.00	.03		.00	.04		−.00	.04
	300	−.50	−.45	.05	.50	.47	.06	−.25	−.24	.05
	500		−.47	.03		.45	.06		−.22	.05
20	300	.00	.00	.04	.00	.00	.03	.00	−.01	.04
	500		.00	.02		.00	.02		.00	.03
	300	−.50	−.47	.03	.50	.45	.05	−.25	−.22	.05
	500		−.47	.02		.48	.05		−.24	.04

Note. True denotes the simulated value, EAP the Expected a Posteriori estimate, and SD the standard deviation.

COMPUTER-BASED ASSESSMENT: MODELING FEEDBACK BEHAVIOR

At a Dutch university of applied sciences, a computer-based assessment of 23 items was used to measure information literacy of 1st-year bachelor students of Law ($N = 218$), Health ($N = 151$), and Business Administration ($N = 241$). The information literacy represents the ability to identify information needs such as locating corresponding information sources, extracting and organizing relevant information from each source, and synthesizing information from different sources (Walraven, Brand-Gruwel, & Boshuizen, 2008). An incomplete test design was used, where the selected Law, Health, and Business Administration students completed 15, 20, and 15 items of the test, respectively. At the start of the assessment, a brief instruction was given and a written instruction was also handed out. Almost every student completed the test within 45 min, and after this time students were allowed to leave the examination room.

After finishing the test, students obtained a knowledge-of-results page showing which items were answered correctly and which were not. Then, students were allowed to use the information retrieval system and to open links to more elaborate feedback pages.

A total of 41% of the students did not consult any of the feedback pages and were classified to the group of nonfeedback users ($G_i = 0$) according to Equation (6). The response data of the other 59% of the students ($G_i = 1$) concerning feedback use were used to estimate the item characteristics and person parameters. The parameters of the joint model were estimated using MCMC, where a run of 15,000 iterations was made, with a burn-in period of 5,000 iterations. Well-known MCMC convergence criteria such as the Gelman-Rubin diagnostic, the Geweke diagnostic for stationarity, the Heidelberger-Welch stationarity and run-length diagnostic, and the Raftery-Lewis run-length diagnostic (all defined in the R-package CODA) showed acceptable results.

The data were not sufficient to support stable item discrimination and time discrimination estimates. Note that around 41% of the students did not consult feedback, which seriously reduced the total amount of data information concerning the time discrimination parameters. Therefore, the discrimination and time intensity parameters were restricted to be common across items. Furthermore, the model was identified by restricting the mean of the item difficulties/thresholds to zero in each measurement model.

Posterior predictive checks for Bayesian item response models were used to evaluate the fit of the different item response model components (e.g., Fox, 2010; Levy, Mislevy, & Sinharay, 2009; Li, Bolt, & Fu, 2006; Patz, Junker, Johnson, & Mariano, 2002; Zhu & Stone, 2011). Comparable posterior predictive checks were used to assess the fit of response time models (e.g., van der Linden, 2006; van der Linden, Klein Entink, & Fox, 2010). When considering the person-fit analysis, a discrepancy measure of replicated sum-scores versus observed sum-scores was used. Only 2 students had an extreme p value, using over 1,000 replicated data sets under the model. One person had a p value of .979, which implies an overprediction by the model. Upon inspection this person had only two items correct. The other person's score was slightly underpredicted. This person scored all items correct and had a p value of .021. The average p value for the response model was .41. This, together with only two extreme observations, indicates a reasonable model fit. For the feedback response model the fit was even better because there were no extreme p values at all, and the average p value was .40. A graphical analysis of model fit for both response time models, using a qq-plot for each item, showed no serious model deviations.

For the joint data analysis with a common covariance structure for the feedback and nonfeedback group, the covariance between ability and speed of working was less than .01. However, when modeling a group-specific covariance structure (feedback and nonfeedback group), a significant positive covariance was found between ability and speed of working for the feedback group ($\sigma_{\theta^a\zeta^a} = .10$) and a significant negative covariance for the nonfeedback group ($\sigma_{\theta^a\zeta^a} = -.11$). In the feedback group, high-performing students worked faster than less performing students. In the nonfeedback group, high-performing students took their time to maximize their results and worked more slowly than less performing students. Due to this difference in strategy, the covariance structure between latent variables was defined to be group specific.

In Table 2, the posterior mean item difficulties (b_k) and time intensities (h_k) are given under the headings "Ability" and "Speed," respectively. Under the headings "Use" and "Attention," the posterior means of the item thresholds (β_k) and time intensities (λ_k) of the feedback behavior assessment are given, respectively. The number of observations and the posterior standard deviations vary across items because an incomplete test design was used. The most difficult items of the information literacy test are related to low threshold values concerning feedback use, which shows that the feedback pages of difficult items are more often consulted than those of easier items. Students most often consulted feedback of items that were answered incorrectly, which were also the more difficult items.

Some of the most difficult items (1, 15, 16, and 18) took apparently on average less time to solve because the corresponding time intensities are relatively low. Often more difficult items require more time to complete (e.g., van der Linden, 2007). However, the corresponding time intensities related to level of attention, λ_k, are relatively high, which indicates that the average attention or processing time was relatively high. On average the difficult items did not appear to require more time to complete, but the corresponding feedback pages were more likely to be consulted and led to higher attention times.

TABLE 2
Item Difficulty and Time Intensity Estimates of the Information Literacy Test, and Item Threshold and Time Intensity Estimates of the Feedback Behavior Assessment

	Information Literacy				Feedback			
	Ability (b_k)		Speed (h_k)		Use (β_k)		Attention (λ_k)	
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	.56	.06	−.27	.27	−.61	.08	−.08	.23
2	−.11	.05	.54	.26	−.33	.08	.25	.23
3	−.26	.05	−.00	.30	.07	.09	.26	.24
4	.02	.05	−.06	.36	−.19	.08	.12	.23
5	−.17	.06	.12	.27	.14	.10	−.07	.26
6	−.03	.05	.17	.27	.20	.09	−.05	.24
7	−.16	.05	.73	.23	.05	.09	.05	.23
8	.10	.05	−.71	.25	−.03	.09	−.11	.24
9	−.24	.07	.72	.34	.30	.10	.17	.26
10	−.48	.05	−.11	.25	.45	.09	.03	.25
11	.16	.05	−.07	.24	.01	.09	.04	.24
12	.07	.05	.18	.25	−.02	.09	.02	.24
13	.10	.05	−.12	.27	−.10	.08	−.26	.24
14	−.35	.06	−.15	.30	.05	.10	−.06	.24
15	.52	.09	−.32	.43	−.54	.12	.12	.25
16	.32	.07	−.35	.28	−.44	.10	.39	.23
17	−.37	.07	−.28	.28	.27	.10	−.24	.25
18	.36	.10	−.13	.47	.05	.13	.14	.27
19	.55	.11	.29	.57	−.18	.13	−.27	.26
20	−.55	.10	.19	.56	.35	.14	−.47	.28
21	.22	.10	−.27	.44	−.10	.13	−.17	.26
22	−.31	.10	−.26	.54	.27	.13	.31	.28
23	.03	.10	.18	.54	.33	.13	−.11	.27

The variability in attention thresholds shows that feedback pages are associated with different processing times, and the variability cannot be directly explained from the associated difficulty or time intensity of the information literacy test items. More research is needed to identify factors explaining differences in attention times and their relationships with test item characteristics.

The estimated population covariance matrix of item parameters given in Table 3, $\mathbf{\Sigma}_I$, shows the variation in item parameters and their covariance structure. The item parameters vary almost equally in difficulty and time intensity. Overall the item parameters do not correlate much because the estimated covariances are around zero. It follows that the highest nonzero covariance $\sigma_{b\beta} = -.07$ (i.e., correlation of –.13) between item difficulty and feedback-use threshold is

TABLE 3
The Estimated Population Covariance Matrix of the Information Literacy Test and Feedback Assessment Parameters

Par.	M	SD
Information literacy, $\boldsymbol{\Sigma}_{I_{11}}$		
τ_b^2	.57	.18
τ_h^2	.72	.24
σ_{bh}	−.02	.14
Feedback, $\boldsymbol{\Sigma}_{I_{22}}$		
τ_β^2	.55	.18
τ_λ^2	.57	.18
$\sigma_{\beta\lambda}$	−.01	.12
Covariance, $\boldsymbol{\Sigma}_{I_{12}}$		
$\sigma_{b\beta}$	−.07	.12
$\sigma_{b\lambda}$	.00	.12
$\sigma_{h\beta}$	.02	.14
$\sigma_{h\lambda}$	.00	.14

TABLE 4
The Estimated Population Means and Covariance Matrix of the Group of Students That Used Feedback and Did Not Use Feedback

		Feedback Use			No Feedback Use			All
Par.		M	SD	Cor	M	SD	Cor	Cor
$\boldsymbol{\mu}_P^a$	μ_{θ^a}	0.04	.03		−.11	.04		
	μ_{ζ^a}	−3.85	.03		−3.67	.04		
$\boldsymbol{\mu}_P^f$	μ_{θ^f}	−.46	.04		–			
	μ_{ζ^f}	−2.21	.04		–			
$\boldsymbol{\Sigma}_{P_{11}}$								
	$\sigma_{\theta^a}^2$	.16	.02	1	.19	.02	1	1
	$\sigma_{\zeta^a}^2$	.19	.02	1	.33	.03	1	1
	$\sigma_{\theta^a\zeta^a}$	.02	.01	.10	−.03	.02	−.11	.01
$\boldsymbol{\Sigma}_{P_{22}}$								
	$\sigma_{\theta^f}^2$	.44	.05	1	0	0	0	1
	$\sigma_{\zeta^f}^2$	.33	.04	1	0	0	0	1
	$\sigma_{\theta^f\zeta^f}$	−.04	.02	−.08	0	0	0	−.04
$\boldsymbol{\Sigma}_{P_{12}}$								
	$\sigma_{\theta^a\theta^f}$	−.04	.02	−.14	0	0	0	−.09
	$\sigma_{\theta^a\zeta^f}$	−.02	.02	−.09	0	0	0	−.08
	$\sigma_{\zeta^a\theta^f}$	−.03	.02	−.10	0	0	0	−.08
	$\sigma_{\zeta^a\zeta^f}$	.04	.02	.15	0	0	0	.10

Note. Par. denotes the parameter, and Cor denotes the correlation.

not significant. In Figure 2, the item difficulties are plotted against the feedback-use thresholds, which shows that the most popular feedback pages (with the lowest thresholds) correspond to the most difficult items.

In Table 4, the estimated mean population parameters are represented for the group of students using feedback and not using feedback. The average item difficulty and time-intensity parameters were fixed to zero. For the feedback groups, the average ability level in information literacy is close to zero, which makes the test suitable for the sample of students. The students of the nonfeedback group score on average lower, with an average value of –.11. The difference in average latent scores is significantly different from zero, where the 95% HPD (Highest Posterior Density) interval equals [.067, .242] and excludes zero. As a result, the average performance of students who consulted at least one feedback page is higher than those who did not consult any of the feedback pages.

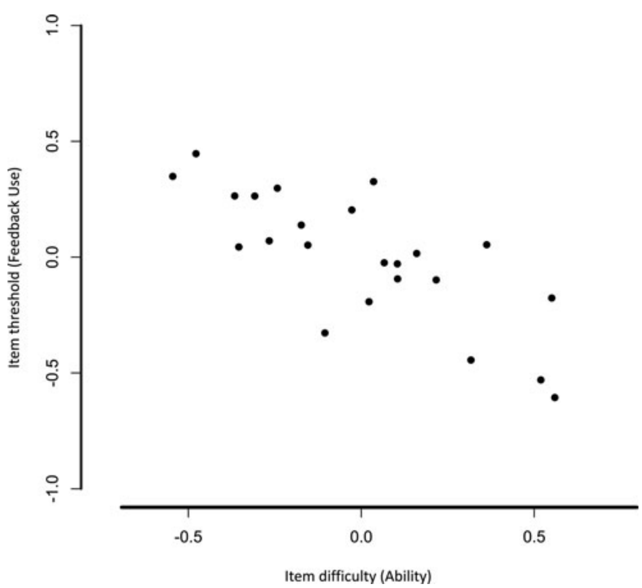

FIGURE 2 Item difficulties of the information literacy test against the item thresholds of using feedback.

The average level of working speed is around –3.85 and –3.67, which means that the expected response time is around $\exp(3.85) = 47$ and $\exp(3.67) = 39$ s for the feedback and nonfeedback group, respectively. For the feedback group, the average propensity to use feedback is –.46, which leads to the average probability of 32% of opening a feedback page. The average level of attention equals –2.21, which leads to an average time of $\exp(2.21) = 9.12$ s of processing item's feedback information.

The population covariance components of the feedback group and the nonfeedback group show that the relationship between ability and speed of working parameter differs over both groups. For the feedback group, there is positive correlation of .10 between ability and speed of working, which means that the more able students work faster. For the nonfeedback group, the relationship is negative with a correlation of –.11, where the able students work more slowly than the less able students. The common correlation estimates under the heading "All" show that the average relation between working speed and ability is close to zero.

In Figure 3, for each student the estimated ability is plotted against the speed of working, where filled circular dots represent students from the feedback group. For the nonfeedback group, it can be seen that many less proficient students worked faster than proficient students. They were also not interested in any of the feedback pages and clearly showed a lack of motivation, which could explain the differences.

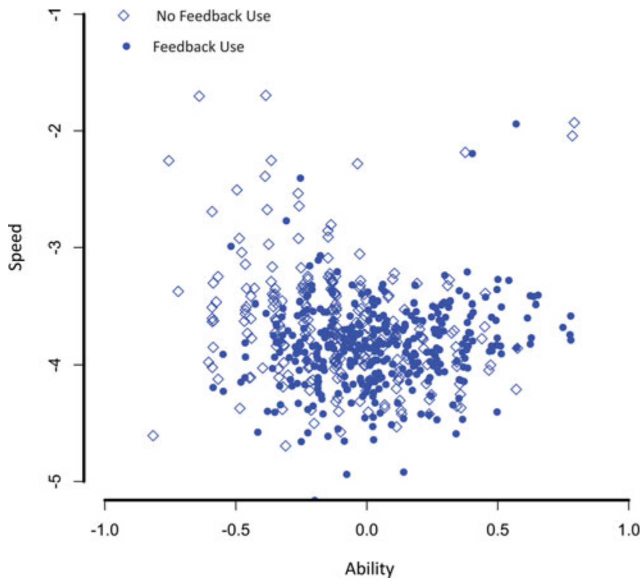

FIGURE 3 Student ability (information literacy) against speed of working for members of the nonfeedback and the feedback group (color figure available online).

In Figure 4, in the left subplot student ability is plotted against the student's propensity of using feedback. The estimated correlation is $\sigma_{\theta^a\theta^f} = -.14$, and a pattern is visible that shows that the more able students were less likely to consult feedback. In fact, the high-ability students did not consult feedback of correctly scored items. This linear relationship was weakened due to the behavior of the less able students, where some of them consulted almost all feedback pages and others consulted only a few pages. That is, some low-ability students requested feedback for all incorrect answers and others for only a few of the incorrect answers. Here, differences in study motivation could also explain the observed differences in feedback behavior of low-ability students.

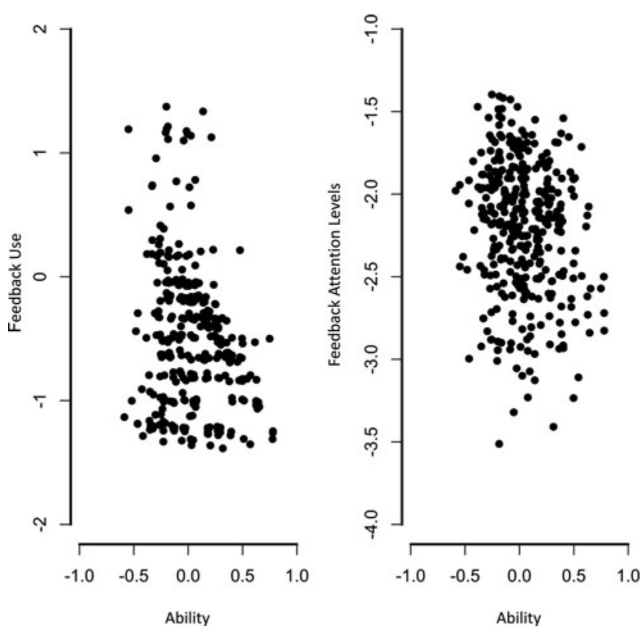

FIGURE 4 Student ability against the student's propensity of using feedback (left plot) and the feedback attention levels (right plot).

The right subplot shows student ability against the level of attention. It can be seen that the reading times differ substantially over students, where high-ability students took slightly more time to process the feedback, $\sigma_{\theta^a\zeta^f} = -.09$. Students' attention correlates negatively with students' propensity to use feedback, $\sigma_{\theta^f\zeta^f} = -.08$. Students who were likely to consult feedback took on average more time to process the feedback information than students who were less likely to consult feedback. Furthermore, the positive correlation $\sigma_{\zeta^a\zeta^f} = .15$ shows that the fast-working students also processed the feedback information faster.

DISCUSSION

With the introduction of computer-based assessments, automatic feedback systems can be developed to improve student learning. Together with tests for learning using a formative assessment, there are opportunities to provide feedback on learning work in a constructive way. This comes with learning and assessment environments that are able to provide student-specific feedback, preferably in an automatic way. New advanced psychometric methods are required to handle associated multiple sources of different types of data on student learning and are capable of measuring student learning and student feedback behavior. In this light, a multivariate hierarchical latent variable model is proposed that provides a way to combine student performance data with feedback behavior data.

The proposed modeling framework can account for the different sources of variation and different types of observed data (i.e., discrete responses and continuous response times). The total variation consists of variation due to the sampling of persons and items and the nesting of responses and times within persons and items. It is also shown that the latent variable approach can handle incomplete test designs. Furthermore, a manifest mixture distribution of the propensity to use feedback can address the subset of students who are not interested in any of the feedback pages.

In the present study, background information was not collected. However, the joint model allows incorporation of student-level variables to explain differences in test performances and feedback behavior. This can be accomplished by modeling the mean structure in Equation (7), which allows multivariate modeling of the latent student variables. In the same way, item-level information can be included in Equation (8) to model differences in item characteristics.

A novel MCMC algorithm is developed to estimate simultaneously all model parameters given unbalanced multivariate mixed response data. At the level of observations, the incomplete test design leads to unbalanced response and response time data, where students complete different subsets of items. At the level of students, although finishing the test, a subset of students did not open any of the feedback pages. This leads to an unbalanced multivariate latent variable structure, which seriously complicates the sampling of the covariance components. The entire covariance matrix cannot be sampled at once and different blocks are sampled in a stepwise way using properties of the multivariate normal distribution. The drawn blocks constitute a new draw of the entire covariance matrix as described in the Appendix.

The MCMC algorithm supports model evaluation via posterior predictive checks. The posterior predictive checks developed by Sinharay, Johnson, & Stern (2006) can be used to check the assumption of local independence. In this context, van der Linden & Glas (2010) proposed lagrange multiplier tests for different local independence assumptions of the response time item response model.

The item characteristics can be informative in different ways for different applications. The construction of informative feedback can be improved when using the item characteristic information. Feedback items with high time-intensive parameters can identify elaborate feedback pages, which are difficult to comprehend but reflect relevant information. These feedback pages can be adjusted to make the displayed material more easily accessible for students. In an adaptive test situation, the automatically generated feedback information can also be adaptive. For example, proficient students may be offered more profound feedback information than less proficient students based on their test results and the characteristics of the feedback items. When test items are constructed according to a rule-based item design (Geerlings, Linden, & Glas, 2013) or by means of item cloning (Geerlings, Glas, & van der Linden, 2011; Glas & van der Linden, 2003), the items may adopt a specific learning strategy, for example. Then, the related feedback information can support the learning strategy and inform students in an adaptive and stepwise way when this is required. One of the goals can be to develop an intelligent tutoring system (Polson & Richardson, 1988), which encourages the students to diagnose their learning status (see also Rupp et al., 2012), to learn by assessment, and to be assisted by an adaptive feedback system. Another application area might be psychological assessment and diagnosis (Templin & Henson, 2006). However, more research is needed to fully control the specification of automatically generated feedback information.

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251–269. doi:10.3102/10769986017003251

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679. doi:10.1080/01621459.1993.10476321

Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive learning environments. *Review of Educational Research*, *73*, 277–320. doi:10.3102/00346543073003277

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–561. doi:10.1007/BF02296195

Bianconcini, S., & Cagnone, S. (2012). A general multivariate latent growth model with applications to student achievement. *Journal of Educational and Behavioral Statistics*, *37*, 339–364. doi:10.3102/1076998610396886

Cool, C., & Belkin, N. (Eds.). (2011). *Interactive information retrieval: History and background*. London, UK: Facet.

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer. doi:10.1007/978-1-4419-0742-4

Geerlings, H., Glas, C., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*, 337–359. doi:10.1007/s11336-011-9204-x

Geerlings, H., van der Linden, W. J., & Glas, C. A. W. (2013). Optimal test design with rule-based item generation. *Applied Psychological Measurement*, *37*, 140–161. doi:10.1177/0146621612468313

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, *27*, 247–261. doi:10.1177/0146621603027004001

Hattie, J., & Timperly, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. doi:10.3102/003465430298487

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48. doi:10.1007/s11336-008-9075-4

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*, 519–537. doi:10.1177/0146621608329504

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3–21. doi:10.1177/0146621605275414

Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, *65*, 391–411. doi:10.1007/BF02296153

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*, 341–384. doi:10.3102/10769986027004341

Polson, M. C., & Richardson, J. J. (Eds.). (1988). *Foundations of Intelligent Tutoring Systems*. New York, NY: Erlbaum.

Ranger, J., & Ortner, T. (2011). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology*, *65*, 334–349. doi:10.1111/j.2044-8317.2011.02032.x

Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S., Crawford, A. V., Calico, T., . . . Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining*, *4*, 49–110.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298–321. doi:10.1177/0146621605285517

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305. doi:10.1037/1082-989X.11.3.287

Timmers, C. F., & Veldkamp, B. P. (2011). Attention paid to feedback provided by a computer-based assessment for learning on information literacy. *Computers & Education*, *56*, 923–930. doi:10.1016/j.compedu.2010.11.007

van der Kleij, F. M., Eggen, T. J. H. M., & Timmers, C. F. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. *CADMO*, *19*, 21–39. doi:10.3280/CAD2011-001004

van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based

assessment for learning. *Computers & Education*, *58*, 263–272. doi:10.1016/jcompedu.2011.07.020
van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204. doi:10.3102/10769986031002181
van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308. doi:10.1007/s11336-066-1478-z
van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247–272. doi:10.1111/j.1745-3984.2009.00080.x
van der Linden, W. J., & Glas, C. A. W. (2010). Statistical tests of conditional independence between responses and response times on test items. *Psychometrika*, *75*, 120–139. doi:10.1007/s11336-009-9129-9
van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347. doi:10.1177/0146621609349800
Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. A. (2008). Information-problem solving: A review of problems students encounter and instructional solutions. *Computers in Human Behavior*, *24*, 623–648.
Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, *66*, 144–168. doi:10.1111/j.2044-8317.2012.02045.x
Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*, 323–339. doi:10.1177/0146621605275984
Zhu, X., & Stone, C. A. (2011). Assessing Fit of Unidimensional Graded Response Models Using Bayesian Methods. *Journal of Educational Measurement*, *48*, 81–97. doi:10.1111/j.1745-3984.2011.00132.x

APPENDIX

Parameter Estimation

The following MCMC scheme is written in R, and the program is available via the authors' website. This MCMC scheme accounts for the mixture distribution of feedback use according to Equation (6). To ease the notation, without explicitly mentioning that $G_i = 1$, the model parameters are sampled given the data of the group that uses feedback (i.e., at least one feedback page is consulted).

RTIRT ability component. The first block of the MCMC algorithm describes the sampling of the RTIRT model parameters. In iteration *m*, the first step is defined as the sampling of the latent explanatory values of ability and speed, denoted as $\theta_i^{a(m)}$ and $\zeta_i^{a(m)}$, respectively. These sampling steps, and the steps for simulating the other RTIRT parameters, are completely described in van der Linden (2007); Klein Entink, Fox, and van der Linden (2009); and Fox (2010). However, the multivariate normal priors for the item parameters and latent variables ability and speed of working are different due to higher level correlation with parameters of the feedback component.

The conditional normal prior $p(\theta_i^a, \zeta_i^a \mid \theta_i^f, \zeta_i^f, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ follows from the multivariate normal prior distribution in Equation (7),

$$\begin{aligned}
&\theta_i^a, \zeta_i^a \mid \theta_i^f, \zeta_i^f, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P \\
&\quad \sim \mathcal{N}\Big(\boldsymbol{\mu}_P^a + \boldsymbol{\Sigma}_{P_{12}}\boldsymbol{\Sigma}_{P_{22}}^{-1}\big((\theta_i^f, \zeta_i^f)^t - \boldsymbol{\mu}_P^f\big), \\
&\quad \boldsymbol{\Sigma}_{P_{11}} - \boldsymbol{\Sigma}_{P_{12}}\boldsymbol{\Sigma}_{P_{22}}^{-1}\boldsymbol{\Sigma}_{P_{21}}\Big). \qquad (10)
\end{aligned}$$

The prior for the item parameters can be deduced from the multivariate normal prior in Equation (8). The multivariate normal prior for the item parameters $\boldsymbol{\xi}_k^a$ given $\boldsymbol{\xi}_k^f$ is given by

$$\begin{aligned}
\boldsymbol{\xi}_k^a \mid \boldsymbol{\xi}_k^f, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I \sim \mathcal{N}\big(&\boldsymbol{\mu}_I^a + \boldsymbol{\Sigma}_{I_{12}}\boldsymbol{\Sigma}_{I_{22}}^{-1}(\boldsymbol{\xi}_k^f - \boldsymbol{\mu}_I^f), \\
&\boldsymbol{\Sigma}_{I_{11}} - \boldsymbol{\Sigma}_{I_{12}}\boldsymbol{\Sigma}_{I_{22}}^{-1}\boldsymbol{\Sigma}_{I_{21}}\big).
\end{aligned}$$

RTIRT feedback component. For the latent variable θ_i^f, the conditional normal prior follows from the multivariate normal prior distribution for the latent person parameters by partitioning the mean $\boldsymbol{\mu}_P$ and covariance matrix $\boldsymbol{\Sigma}_P$ accordingly. Let $\mu_{\theta f}$ and $\sigma_{\theta f}$ denote the prior mean and variance parameter of the conditional normal prior distribution for θ_i^f. The conditional posterior distribution is normal,

$$\begin{aligned}
&\theta_i^f \mid \mathbf{z}_i^f, \boldsymbol{\Sigma}_P, \boldsymbol{\mu}_P, \theta_i^a, \zeta_i^a, \zeta_i^f, \boldsymbol{\xi}^f \\
&\quad \sim N\big(E(\theta_i^f \mid z_i^f), Var(\theta_i^f \mid \mathbf{z}_i^f)\big),
\end{aligned}$$

with

$$\begin{aligned}
E\big(\theta_i^f \mid \mathbf{z}_i^f, \boldsymbol{\Sigma}_P, \boldsymbol{\mu}_P\big) &= \big((\boldsymbol{\alpha}^t\boldsymbol{\alpha})^{-1} + \sigma_{\theta f}^{-1}\big)^{-1} \\
&\quad \times \Big(\boldsymbol{\alpha}^t(\mathbf{z}_i^f + \boldsymbol{\beta}) + \boldsymbol{\mu}_{\theta f}/\sigma_{\theta f}\Big)
\end{aligned}$$

and

$$Var\big(\theta_i^f \mid \mathbf{z}_i^f, \boldsymbol{\Sigma}_P, \boldsymbol{\alpha}\big) = \big(1/\boldsymbol{\alpha}^t\boldsymbol{\alpha} + 1/\sigma_{\theta f}\big)^{-1}.$$

For the latent variable ζ_i^f, the conditional normal prior also follows from the multivariate normal prior distribution for the latent person parameters. Partition the mean $\boldsymbol{\mu}_P$ and covariance matrix $\boldsymbol{\Sigma}_P$ and let $\mu_{\zeta f}$ and $\sigma_{\zeta f}$ denote the prior mean and variance parameter of the conditional normal prior distribution for ζ_i^f. It follows that the conditional posterior distribution is normal with parameters

$$\begin{aligned}
&E\big(\zeta_i^f \mid \mathbf{y}_i^f, \mathbf{t}_i^f, \boldsymbol{\Sigma}_P, \boldsymbol{\mu}_P, \boldsymbol{\xi}^f\big) \\
&\quad = \left(\sum_{k|i} d_{ik}/\phi_k^2 + \sigma_{\zeta f}^{-1}\right)^{-1} \\
&\quad \times \left(\sum_{k|i} d_{ik}\phi_k(t_{ik}^f + \lambda_k) + \boldsymbol{\mu}_{\zeta f}/\sigma_{\zeta f}\right)
\end{aligned}$$

and

$$Var\big(\zeta_i^f \mid \mathbf{t}_i^f, \boldsymbol{\Sigma}_P, \boldsymbol{\phi}\big) = \left(\sum_{k|i} d_{ik}/\phi_k^2 + 1/\sigma_{\zeta f}\right)^{-1}.$$

Note that the attention is conditionally modeled given feedback use where $d_{ik} = 1$ when $Y_{ik}^{f} = 1$ and zero otherwise.

The prior for the item parameters of the feedback component can be deduced from the multivariate normal prior. Therefore,

$$\boldsymbol{\xi}_k^f \mid \boldsymbol{\xi}_k^a, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I \sim \mathcal{N}\big(\boldsymbol{\mu}_I^f + \boldsymbol{\Sigma}_{I_{21}} \boldsymbol{\Sigma}_{I_{11}}^{-1} (\boldsymbol{\xi}_k^a - \boldsymbol{\mu}_I^a), \\ \boldsymbol{\Sigma}_{I_{22}} - \boldsymbol{\Sigma}_{I_{21}} \boldsymbol{\Sigma}_{I_{11}}^{-1} \boldsymbol{\Sigma}_{I_{12}}\big),$$

Another partitioning can be made to define the conditional prior for the item parameters (α_k, β_k) and (ϕ_k, λ_k). The item parameters (α_k, β_k) are the coefficients of the regression of z_k^f on $\mathbf{H}_I = (\boldsymbol{\theta}^f, -\mathbf{1}_N)$. It follows that this leads to a multivariate normal conditional posterior distribution for the item parameters (α_k, β_k).

In the same way, the parameters (ϕ_k, λ_k) are the coefficients of the regression of $\boldsymbol{T}_k^f$ on $\mathbf{H}_I = (\boldsymbol{\zeta}^{f*}, -\mathbf{1}_{N_k^f})$, where N_k^f denotes the number of participants consulting the feedback to item k and $\boldsymbol{\zeta}^{f*}$ the corresponding vector storing their level of attention. This leads to a multivariate normal conditional posterior distribution for the item parameters (ϕ_k, λ_k).

Sampling of covariance and mean population parameters. Sampling covariance parameters, $\boldsymbol{\Sigma}_P$, is carried out in three sampling steps. Let $\boldsymbol{\Omega}^a = (\boldsymbol{\theta}^a, \boldsymbol{\zeta}^a)$ and $\boldsymbol{\Sigma}_{P_{11}}$ has an inverse Wishart prior with N_0 degrees of freedom and scale parameter $\boldsymbol{\Sigma}_0$. The conditional posterior is defined as

$$\boldsymbol{\Sigma}_{P_{11}} \mid \boldsymbol{\Omega}^a, \boldsymbol{\mu}_P \sim inverse-Wishart\,(N + N_0, \mathbf{S}), \quad (11)$$

where $\mathbf{S} = \sum_i (\boldsymbol{\Omega}_i^a - \boldsymbol{\mu}_P)(\boldsymbol{\Omega}_i^a - \boldsymbol{\mu}_P)^t + \boldsymbol{\Sigma}_0$.

The component $\boldsymbol{\Sigma}_{P_{12}}$ is sampled as the regression parameter in the regression of $\boldsymbol{\Omega}_i^a$ on $\boldsymbol{\Omega}_i^f$ according to Equation (10) given $\boldsymbol{\Sigma}_{P_{11}}$. A vague proper prior is defined for the covariance component $\boldsymbol{\Sigma}_{P_{12}}$.

The component $\boldsymbol{\Sigma}_{P_{22}} - \boldsymbol{\Sigma}_{P_{21}} \boldsymbol{\Sigma}_{P_{11}}^{-1} \boldsymbol{\Sigma}_{P_{12}}$ is the covariance matrix of $\boldsymbol{\Omega}_i^f$ given $\boldsymbol{\Omega}_i^a$. A sample is drawn from the inverse Wishart distribution with $N + N_0$ degrees of freedom and sale parameter

$$\mathbf{S} = \sum_i \left(\boldsymbol{\Omega}_i^f - \left(\boldsymbol{\mu}_P^f + \boldsymbol{\Sigma}_{P_{21}} \boldsymbol{\Sigma}_{P_{11}}^{-1} \left(\boldsymbol{\Omega}_i^a - \boldsymbol{\mu}_P^a\right)\right)\right) \\ \times \left(\boldsymbol{\Omega}_i^f - \left(\boldsymbol{\mu}_P^f + \boldsymbol{\Sigma}_{P_{21}} \boldsymbol{\Sigma}_{P_{11}}^{-1} \left(\boldsymbol{\Omega}_i^a - \boldsymbol{\mu}_P^a\right)\right)\right)^t.$$

The draws of $\boldsymbol{\Sigma}_{P_{12}}$ and $\boldsymbol{\Sigma}_{P_{11}}$ are used to compute $\boldsymbol{\Sigma}_{P_{21}} \boldsymbol{\Sigma}_{P_{11}}^{-1} \boldsymbol{\Sigma}_{P_{12}}$, which is used to obtain a sample of the component $\boldsymbol{\Sigma}_{P_{22}}$.

The covariance parameters of the item parameters, $\boldsymbol{\Sigma}_I$, are sampled in three comparable steps. The covariance parameters of $\boldsymbol{\xi}^a$ are sampled in the first step. In the second step, the covariance parameters $\boldsymbol{\Sigma}_{I_{22}} - \boldsymbol{\Sigma}_{I_{21}} \boldsymbol{\Sigma}_{I_{11}}^{-1} \boldsymbol{\Sigma}_{I_{12}}$ are sampled. Finally, the covariance parameters, $\boldsymbol{\Sigma}_{I_{12}}$, are sampled that specify the covariance structure between $\boldsymbol{\xi}^a$ and $\boldsymbol{\xi}^f$. The posterior distributions can be derived in a way similar to the covariance parameters of the person parameters.

The mean population parameters for item and person parameters, $(\boldsymbol{\mu}_I, \boldsymbol{\mu}_P)$, are sampled from multivariate normal distributions. Using a multivariate normal prior with mean $\boldsymbol{\mu}_0$ and variance parameter $\boldsymbol{\Sigma}_0$, the conditional posterior distribution of $\boldsymbol{\mu}_P$ is given by

$$\boldsymbol{\mu}_P \mid \mathbf{z}, \mathbf{t}, \boldsymbol{\Omega}_P, \boldsymbol{\Sigma}_P \\ \sim \mathcal{N}\left(\frac{\overline{\boldsymbol{\Omega}}_P N \boldsymbol{\Sigma}_P^{-1} + \boldsymbol{\mu}_0 \boldsymbol{\Sigma}_0^{-1}}{N \boldsymbol{\Sigma}_P^{-1} + \boldsymbol{\Sigma}_0^{-1}}, \frac{1}{\boldsymbol{\Sigma}_0^{-1} + N \boldsymbol{\Sigma}_P^{-1}}\right), \quad (12)$$

where $\overline{\boldsymbol{\Omega}}_P$ represents the vector of average latent variable values across persons.

In the same way, the conditional posterior distribution of $\boldsymbol{\mu}_I$ is given by

$$\boldsymbol{\mu}_I \mid \mathbf{z}, \mathbf{t}, \boldsymbol{\Omega}_I, \boldsymbol{\Sigma}_I \\ \sim \mathcal{N}\left(\frac{\overline{\boldsymbol{\Omega}}_I K \boldsymbol{\Sigma}_I^{-1} + \boldsymbol{\mu}_0 \boldsymbol{\Sigma}_0^{-1}}{K \boldsymbol{\Sigma}_I^{-1} + \boldsymbol{\Sigma}_0^{-1}}, \frac{1}{\boldsymbol{\Sigma}_0^{-1} + K \boldsymbol{\Sigma}_I^{-1}}\right), \quad (13)$$

where $\overline{\boldsymbol{\Omega}}_I$ represents the vector of average item values across items.