Journal of Computerized Adaptive Testing

Volume 5 Number 2

April 2017

Latent-Class-Based Item Selection for Computerized Adaptive Progress Tests

Nikky van Buuren and Theo H. J. M. Eggen

DOI 10.7333/1704-0502022

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 5, Number 2, April 2017

DOI 10.7333/1704-0502022

ISSN 2165-6592

Latent-Class-Based Item Selection for Computerized Adaptive Progress Tests

Nikky van Buuren and Theo H. J. M. Eggen

University of Twente and Cito

Standard computerized adaptive testing (CAT) methods require an underlying item response theory (IRT) model. An item bank can be constructed from the IRT model, and subsequent items can be selected with maximum information at the examinee's estimated ability level. IRT models, however, do not always fit test data exactly. In such situations, it is not possible to employ standard CAT methods without violating assumptions. To extend the scope of adaptive testing, this research shows how latent class analysis (LCA) can be used in item bank construction. In addition, the research investigates suitable item selection algorithms using Kullback-Leibler (KL) information for item banks based on LCA. The KL information values can be used to select items and to construct an adaptive test. Simulations show that item selection based on KL information outperformed random selection of items in progress testing. The effectiveness of the selection algorithm is evaluated, and a possible scoring for the new adaptive item selection with two classes is proposed. The applicability of the methods is illustrated by constructing a computerized adaptive progress test (CAPT) on an example data set drawn from the Dutch Medical Progress Test.

Keywords: *Latent class analysis, computerized adaptive progress test, Kullback-Leibler information, item selection method, log-odds scoring*

Three aspects of using latent class analysis (LCA) in computerized adaptive tests (CATs) are addressed in this research: item bank construction, item selection, and a possible scoring procedure. Each is illustrated with an example data set drawn from the Dutch Medical Progress Test. The research investigates the feasibility of developing a CAT using items that do not correspond exactly to an item response theory (IRT) framework when applying the proposed latent-class-based algorithms for CAT. The selection algorithm makes use of Kullback-Leibler (KL) selection methods, which have previously been applied for other purposes such as CATs for cognitive diagnosis (Cheng, 2009).

Standard CAT

CAT normally refers to methods of testing in which each subsequent item is the item with the maximum Fisher information for examinee's estimated latent ability, $\hat{\theta}$. This means that each examinee takes a tailored version of a test drawn from a certain bank of items. The development and application of CAT has increased over recent decades, and several educational institutions and testing centers have integrated adaptive methods into their tests (Weiss & Kingsbury, 1984; Van der Linden & Veldkamp, 2004; Thompson & Weiss, 2011). CAT has many advantages for both examinees and evaluators. These advantages are mainly expressed in terms of efficiency and a better user test experience, as the questions are tailored to an individual's ability level (Eggen, 2008).

Misfit in IRT Models

In some situations, however, the standard IRT framework that underlies CAT does not fit the data, for example, during the calibration phase when building a CAT. Global fit of the IRT models to an item bank can be tested by looking at global fit measures such as the Q_1 test, R_1 test, and/or likelihood ratio tests (Andersen, 1973; Suárez-Falcón & Glas, 2003). These testing methods for global model fit can be first pointers to misfit. Concerns raised by global indications of misfit can be caused by violations of one or more assumptions of the IRT model. Item response functions can be flat instead of S-shaped, and problems with assumed local independence and unidimensionality can occur (Yang & Kao, 2014).

If many items from the item bank show misfit, constructing an item bank might be impossible or only possible with heavy violations of the assumptions of standard IRT models. In these cases, other models can be applied to the data, such as multidimensional IRT models or latent variable mixture models. Multidimensional IRT can be modeled when multiple constructs in the data disturb the model fit (Béguin & Glas, 2001). Latent variable mixture modeling is useful in populations with heterogenous rather than homogenous samples (Sawatzky, Ratner, Kopec, Wu, & Zumbo, 2016). Sometimes, however, these other IRT models do not fit

specific types of items. This research, therefore, explores the opportunities for applying and adapting current methods and models for CAT if the existing alternatives to standard IRT in CAT do not work.

Research on LCA and CAT

The present research appears to be the first to propose LCA in combination with CAT in an actual testing situation. Macready and Dayton (1992) investigated the use of LCA in CAT, concluding that the combination of CAT with LCA allows for conceptually simpler models than CAT with IRT. Macready and Dayton also acknowledged that LCA has fewer untestable assumptions. Their goal was to classify respondents into the appropriate latent class with CAT and to obtain an acceptable level of error.

Later, Cheng (2009) used LCA models and proposed an item selection method for CAT based on KL information, which was used in this research. However, the focus of Cheng's application was on cognitive diagnostic computerized adaptive testing (CD-CAT). This method aims to classify examinees into latent classes, whereas the progress tests aim to compare students within and across certain classes after being classified in a latent class.

Xu, Chang, and Douglas (2003) were the first to propose item selection algorithms for CD-CAT, and Cheng has extended these methods. Other research on CD-CAT is that of Wang, Chang, and Douglas (2012). They identified an effective item selection algorithm for CAT that not only efficiently estimates θ but also classifies the student's knowledge status based on a Q-matrix. The KL item selection method, which has been used in other research (Chen & Ankenman, 2004; Mulder & Van der Linden, 2009; Eggen, 2012), would probably be highly valuable in situations where standard IRT models do not fit the data. To determine whether the method is truly beneficial, these methods were tested in the present research by constructing and evaluating a computerized adaptive progress test (CAPT).

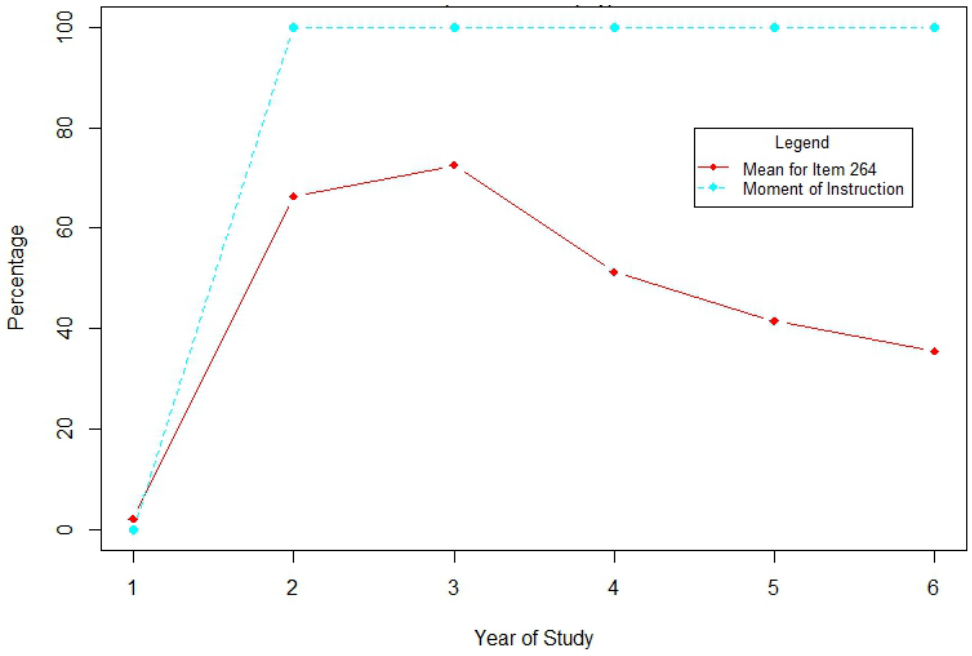
Progress Data and Misfit

Progress tests are longitudinal educational assessments that are intended to measure learning outcomes over the course of learning processes (Tio, Schutte, Meiboom, Greidanus, Dubois, Bremers, & the Dutch Working Group of the Interuniversity Progress Test of Medicine, 2016). Medical progress test data from Dutch universities are used as an example throughout this paper. Progress testing usually involves a large number of items, which are administered to all students irrespective of their stage in the curriculum. In the medical progress test, the students in the beginning stage of the curriculum are typically able to answer only a few items correctly (Wrigley, Van der Vleuten, Freeman, & Muijtens, 2012). Therefore, the application

of a CAT framework to items in a progress test could be beneficial in terms of efficiency and improved experience for students.

Standard IRT models do not always fit the data of progress tests. In the calibration phase, it was difficult to fit all items onto one IRT scale because global fit indices, as well as many item-specific indicators, showed misfit. Upon closer inspection, it appeared that a large number of the items behaved unexpectedly by showing a "jump" in probability at a certain point in time. When looking at the instruction program of one university, this jump usually coincided with the moment of instruction of the topic. Following this moment of instruction, the knowledge of all students immediately changed; however, this was not necessarily associated with their θ level. In addition, it appears that for most of these items, the knowledge that had been gathered did not always grow in the remaining years of study. An example of one of the items with misfit can be seen in Figure 1.

Figure 1. Mean Percentage of Correct Responses on One Jump-Type Item from the 2005 Dutch Medical Progress Test Over the Study Years



A mix of two item types can occur in progress test data: (1) jump-type items and (2) items that can be scaled correctly on an IRT scale without misfit. This can lead to a mix of items that become easier for students with higher ability levels and items that are easier or more difficult at a certain point in time due to the moment of education. The former type of items generally fit into an IRT framework, whereas the latter type disturbs the general assumption of higher

ability associated with a higher probability of answering items correctly. The moments of instruction differ over universities, causing the order of difficulty of items to change over universities and study years. The assumption of constructing a population model is not only violated in multidimensional IRT models but also in latent variable mixture models. Introducing multiple constructs and correcting for heterogeneous samples do not overcome this violation.

The standard IRT-based methods can show misfit for progress test data and are not applicable in all testing situations. Therefore, this paper demonstrates a latent-class-based item selection method for progress test items and evaluates the performance of this selection method and algorithm in CAPT simulations. The item selection in the test is based on KL information. An illustrated example of the selection algorithm is given by building an item bank and performing CAPT simulations for the Dutch Medical Progress Test, which has been a standard paper-and-pencil test for many years (Schuwirth, Bosman, Henning, Rinkel, & Wenink, 2010). Finally, this paper proposes and evaluates a log-odds scoring method based on posterior class probabilities.

Background

Item Response Theory and CAT

Typically, methods used to construct a CAT first fit a model from IRT to the data in order to construct a calibrated item bank (Embretson & Reise, 2000). Using the estimated parameters from these IRT models, tailored item selection is possible for each examinee. Two examples of applying IRT models to a set of items are fitting a one-parameter logistic model, which was first proposed by Rasch (1960), or fitting a two-parameter logistic model (Wainer, 2000).

After item bank calibration, an item selection algorithm is applied to continue the construction of the adaptive test. A strategy for selecting the next item in a test, given an estimate of ability, $\widehat{\theta}_i$, and based on the previous responses of examinee *i*, is a method that selects the item providing maximum information (Van der Linden & Glas, 2000). This method chooses an item *j**, with possible values $j = 1, \ldots, J$, which maximizes the item information at the estimated ability level $\widehat{\theta}_i$ of the examinee, as follows:

$$Ij(\widehat{\theta}_i) = \frac{[P'_j(\widehat{\theta}_i)]^2}{\{P_{j'}(\widehat{\theta}_i)[1 - P_{j'}(\widehat{\theta}_i)]\}}. \quad (1)$$

In this equation, $P_{j'}(\widehat{\theta}_i)$ is the probability of a correct response to item *j* from the calibrated item bank for examinee i at a current $\widehat{\theta}_i$. Each computerized adaptive test also needs a stop-

ping rule. For example, the adaptive test could be terminated after a fixed number of items or when a certain precision level has been reached (Wang, Chang, & Boughton, 2013). The decision on a stopping rule depends on many factors and can result in either a fixed-length or variable-length test.

The Dutch Medical Progress Test

In five of the eight medical schools in the Netherlands, a medical progress test is administered quarterly for all students. All first- to sixth-year students answer the same 200 items on various medical topics in a paper-and-pencil format. According to Schuwirth et al. (2010), "the longitudinal feature of the progress test provides a unique and demonstrable measurement of the growth and effectiveness of students' knowledge acquisition." The goal of the test is to monitor the progress of students throughout the course of the medical program. Data from the Dutch Medical Progress Test over seven subsequent years (2005 to 2011) were used in this research to illustrate the use of the proposed latent-class-based item selection algorithm and its efficacy in simulations of the CAPT.

Item Bank

The data from the Dutch Medical Progress Test used for this research consisted of responses from all the medical students at five universities. The sample involved the December versions of the tests from 2005 to 2011. Consequently, a total of 1,400 different items were available for possible selection in the item bank for CAPT. None of the 1,400 items occurred more than once in the tests. Over the years 2005 through 2011, the sample sizes differed between a minimum of N = 5,318 students in 2005 and a maximum of N = 7,001 in 2008. It is assumed that students participating in the December tests had the same ability distribution over the years, meaning that new respondents would be acceptable substitutes for respondents leaving the sample.

Data analyses provided relevant information regarding performance on the items for students in different years of their medical studies. Over the study year groups, the probability of answering an item correctly increased gradually for some items, while for other items a clear jump could be observed in subsequent years. This jump could have many causes. For instance, the moment at which students have been taught and/or the origin of the knowledge or skill tested by the item could be of influence. A "jump-type item" was defined as an item that showed an increase of at least .30 in the mean probability of being answered correctly over two subsequent study year groups. For example, the growth in mean probability must be observed between Study Year 1 and 2 or between Study Year 4 and 5. In the data set, 208 of the 1,400 item total could be classified as jump-type items.

"Grow-type" items seemed to follow an IRT model in item bank calibrations. Jump-type items, however, violated some of the IRT assumptions, such as monotonicity and invariance of item parameters and the latent ability across different universities and study years. It was hypothesized that LCA could be an alternative to IRT in this case as well as in other cases where IRT does not fit. LCA provides a classification of students instead of a θ estimate on a continuous scale; this classification can also provide valuable information in addition to an IRT model or as a substitute for a continuous outcome.

Using LCA Outcomes

IRT models estimate the ability of the student on a continuous latent scale. This research proposes another option, namely, the path of classification and probabilistic information on class memberships. This means that in constructing an item bank with LCA, students are first clustered into a certain number of classes, which are restricted as follows: The best students are in Class 1, and the least able students are in the highest class, depending of the number of classes in the solution. Being a member of a certain latent class corresponds to a certain probability of answering an item correctly.

This probabilistic information can be used to calculate the latent class membership probabilities for each class and can be highly informative. In practice, there are three options for using the outcome of a CAT based on LCA when constructing tests. First, the latent class membership outcome can simply be used as a classification. Second, classification information can be used in addition to a standard $\hat{\theta}$ estimation coming from IRT. Finally, the probabilistic outcomes on the class membership can be used to provide information about the ability of the student.

Latent Class Analysis

LCA is a method of analysis in which observable or manifest variables are related to unobservable or latent variables (Lazarsfeld & Henry, 1968). Both the observed and unobserved variables in this model are assumed to be categorical. This attribute is the essential difference when compared to psychological measurement models such as IRT. In IRT, the observed variables would also be categorical, but the underlying latent variable would be continuous. Essentially, LCA models provide a probabilistic or fuzzy outcome with respect to predicted class membership, whereas IRT models give a scaled latent estimate on a continuous trait.

In a latent class model, examinees within the same latent class have common characteristics with regard to certain criteria, and examinees in different latent classes are dissimilar from each other. In LCA, parameters are estimated for class profiles and the size of each class. For multiple dichotomous items, the predicted latent class memberships $\mathrm{P}(C = c|\mathbf{x}_i)$ can be cal-

culated given the observed answer pattern $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})$ of an examinee i (Dayton, 1998). This holds for dichotomous items, where $x_{ij} \in \{0,1\}$, and $i = 1, \ldots, N$ for the persons and $j = 1, .., J$ for the items. A latent class is notated as C, with $c = 1, \ldots,$ Tc, with Tc being the total number of latent classes. The joint probability of obtaining a certain response pattern $\mathbf{x}_i$ would then be

$$\mathrm{P}(\mathbf{x}_i) = \sum_{c=1}^{Tc} \pi_c \prod_{j=1}^{J} \mathrm{P}(x_{ij} = 1 | C = c), \tag{2}$$

where π_c indicates the proportion of students that belong to class *C*. The assumption of local independence can be recognized in Equation 2 because the *J* observed items are assumed to be mutually independent within each class *C*. An interesting aspect of this research and the use of LCA is that the model solutions assign students to a latent class with a certain probability. The estimated class probability for each individual to belong to a class *C*, given person *i*'s answer pattern $\mathbf{x}_i$, is

$$\mathrm{P}(C = c | \mathbf{x}_i) = \frac{\mathrm{P}(\pi_c)\mathrm{P}(\mathbf{x}_i | C = c)}{\mathrm{P}(\mathbf{x}_i)}, \tag{3}$$

which is referred to as a posterior probability. For example, a correct answer to an item for a person would result in the following expression:

$$\mathrm{P}(C = c | x_{ij} = 1) = \frac{\mathrm{P}(x_{ij} = 1 | C = c)\pi_c}{\mathrm{P}(x_{ij} = 1)}. \tag{4}$$

In LCA, each student with an observed response pattern $\mathbf{x}_i$ would belong to each of the classes with a certain probability. When the goal of the latent class analysis is person classification, as is the case in this application, the most commonly used classification rule is modal assignment. This rule (Hagenaars & McCutcheon, 2002) dictates that the examinee will be assigned to the class with the highest $\mathrm{P}(C = c | \mathbf{x}_i)$. Restrictions can be applied to LCAs and may be necessary when constructing an item bank with items from non-overlapping examinees.

Latent Gold 4.5 software (Vermunt & Magidson, 2000; Haughton, Legrand, & Woolford, 2009) was used for the LCA analyses in this research. Several indices can be considered in drawing conclusions about how many classes are required to give the best fit to the data and to

determine optimum class sizes. One index measure and global fit index is the -2 log likelihood (-2LL), which compares several models with varying numbers of classes (Nylund, Asparouhov, & Muthén, 2008). Another fit index is the Bayesian information criterion (BIC; Schwartz, 1978), which is regarded as a good indicator for distinguishing between models with varying numbers of latent classes (McCutcheon, 1987; Vermunt & Magidson, 2000; Hagenaars & McCutcheon, 2002). The BIC is defined as

$$BIC = -2\log L + p\log(N), \tag{5}$$

where p is the number of free model parameters and N represents the sample size. This information criterion also uses -2LL but adjusts for the number of parameters and the sample size in the model, resulting in improved index reliability (Nylund et al., 2008).

Restrictions

The data from the Dutch Medical Progress Test are not linked, which means students cannot be identified over the years. The unlinked design implies a few restrictions when constructing the item bank by applying LCA. The linking of the data and fitting latent class models for each year can be done by determining the optimal number of latent classes over the years and fitting LCA models to them. Next, a group-specific latent class model can be fitted by imposing a restriction on class sizes, which means that class size in each latent class model over the years must be equal. This model then has the following notation: Group refers to the year of the test and is indicated by $g \in \{1, \ldots, G\}$, and an item is indicated by $j \in \{1, \ldots, J_g\}$, with J_g being the total number of items in group g. This leads to the following extension of the joint probability formula:

$$\mathrm{P}(x_{i1}^{(g)}, \ldots, x_{iN}^{(g)}) = \sum_{c=1}^{Tc} \pi_c^g \prod_{j=1}^{J_g} \mathrm{P}\left(x_{ij}^g = 1 \middle| C = c\right). \tag{6}$$

When separate models have been estimated and the optimal number of latent classes has been determined, the mean $\bar{P}_c$ can be found by calculating the average class size over all groups g with

$$\bar{P}_c = \frac{1}{G}\sum_{c=1}^{Tc} \pi_c^g \tag{7}$$

for all latent classes $\in \{1, \ldots, Tc\}$.

Kullback–Leibler Selection Algorithm

The proposed item selection algorithm for item banks based on LCA makes use of KL information. The general form of KL information (Kullback & Leibler, 1951; Cover & Thomas, 1991) between two probability distributions is expressed as

$$D[f,g] = E_f\left[\log\frac{f(\mathbf{x})}{g(\mathbf{x})}\right]. \tag{8}$$

The probability distribution $f(\mathbf{x})$ would usually represent the "true" distribution of the data, or a precise theoretical distribution. Normally, $g(\mathbf{x})$ would be the representation of a model or an approximation of $f(\mathbf{x})$. Therefore, it reflects a distance or divergence between these two distributions.

Note that it is not a distance in mathematical terms, as the measure is not symmetric with $D[f,g] \neq D[g,f]$. Large KL information values of $D[f,g]$ would be an indication of two statistically different distributions, indicating deviances between the two (Cheng, 2009). Although several terms are used in the literature—such as KL distance, KL divergence, and KL information—the latter term is used here.

Cheng (2009) has proposed an application of the KL algorithm that helps select items in order to improve cognitive diagnosis. The focus on cognitive diagnosis is to classify persons into cognitive profiles, a_i. The true state of a person's attribute profile is unknown. Therefore, the KL information of the conditional distribution of response pattern $\mathbf{x}_i$, given the current estimate of person i's latent cognitive diagnosis state and the conditional distribution of $\mathbf{x}_i$, given other latent cognitive diagnosis states, can be measured. When applying KL information, it can be calculated between $f(\mathbf{x}_i|\,\widehat{a_i^{(t)}})$, with $\widehat{a_i^{(t)}}$ indicating the current estimate of $\widehat{a_i}$, and another distribution based on $\mathbf{x}_i$, given another latent state a_c:

$$D_j(\widehat{a_i^{(t)}}\,\|\,a_c) = \sum_{q=0}^{1}\log\left(\frac{\mathrm{P}(\mathrm{x}_{ij}=q\,|\,\widehat{a_i^{(t)}})}{\mathrm{P}(\mathrm{x}_{ij}=q\,|\,a_c}\right)\mathrm{P}\left(\mathrm{x}_{ij}=q|\widehat{a_i^{(t)}}\right). \tag{9}$$

Equation 9 can be used for models with two latent states. However, if more than one alternative latent state is available, the sum of the KL information between the estimated current latent state $\widehat{a_i^{(t)}}$ and all other latent states a_c could be used (Xu, Chang, & Douglas, 2003):

$$KL_j\left(\widehat{a_i^{(t)}}\right) = \sum_{c=1}^{C} D_j(\widehat{a_i^{(t)}}\,\|\,a_c). \tag{10}$$

Item selection would then follow, based on the aim of finding the item with the maximum $KL_j\left(\widehat{a_i^{(t)}}\right)$ for an examinee in a specific latent state. Based on the maximum KL information of an item, given the current estimate of the latent state, the $(t+1)^{th}$ item will be chosen. If there are two latent states, summing the KL information using Equation 10 is not necessary.

Cheng (2009) also proposed extensions of the KL information described above, one of which is posterior weighted KL information (PWKL). The PWKL could incorporate information about old samples into the analysis of current samples with the help of prior information.

Simulations of CAPTs

Method

In each simulated CAPT, 150 examinees per latent class were assigned to their "true" latent state. Thus, the two-class LCA solution simulation tested the efficacy of reclassifying 300 examinees based on the simulated response patterns, whereas the three-class solution tested the classification accuracy for 450 examinees. Item responses were generated based on item response probabilities, given the "true" class to which the examinee belonged. Replication of these simulations, 50 per setting, made it possible to draw conclusions about variances in various test situations.

A random generator from the uniform distribution with $d \sim \mathrm{U}(0,1)$ was used to simulate these answer patterns. A total of D uniformly distributed numbers were drawn from this distribution for each of the 50 replications. The 208 jump-type items in the databank and the LCA two-class and three-class solutions based on these items had the estimated probabilities $\mathrm{P}(x_{ij}=1|C=c)$. To generate answer patterns, the probabilities $\mathrm{P}\left(x_{ij}=1\middle|C=c\right)$, as calibrated from the LCA, were compared to the boundary values d,

$$\text{where } x_{ij}\begin{cases}0 \text{ for } \mathrm{P}\left(x_{ij}=1\middle|C=c\right) \geq d_{ij}\\ 1 \text{ for } \mathrm{P}\left(x_{ij}=1\middle|C=c\right) < d_{ij}\end{cases}.$$

Response patterns conditioned on latent class membership were generated in the statistical program R (version 3.2.2 for Windows).

The simulated answer patterns were then again used to estimate the latent class membership. In contrast this time only, a subset of the item responses were used to estimate class memberships. The idea of the CAPT is to use fewer items than in the current 200-item progress test and correctly predict the true latent class and latent ability of the examinees. Two methods of item selection were compared: (1) the baseline, which involved random selection of the items, and (2) the experimental setting, which was expected to increase the performance of the CAPT and in which the selection was based on KL information.

For each examinee, the random selection method generated a random vector of test length 10, 15, or 20 out of the 208 jump-type items. The CAT selection method based on KL information also generated tests of length 10, 15, and 20 for each examinee; but the items were chosen based upon the current class estimate $a_i^{(t)}$. After each question, the current estimate of class a_i^t was used to choose the best-fitting item according to the KL information indices.

The predicted latent class probabilities were updated after an item was administered; and the next item was then chosen by applying the modal classification rule, based on the largest KL information corresponding to the predicted latent state a_i^t. In contrast to the CAT with an underlying item response model, where the ability parameter on a continuous latent scale was used to select the optimal next item, the CAPT for the jump-type items used the categorical class membership estimate to select the next item. The items with the maximum amount of KL information contributed most to the correct classification of the examinees.

Results

Calibrating an item bank with LCA. Latent class models with varying number of classes were fitted to the medical progress test data in order to form classes of students. Table 1 shows that the BIC of the four-class solution had the lowest value and seemed to fit the data best according to the fit indices. However, other fit measures also were examined in deciding on the best usable LCA solution.

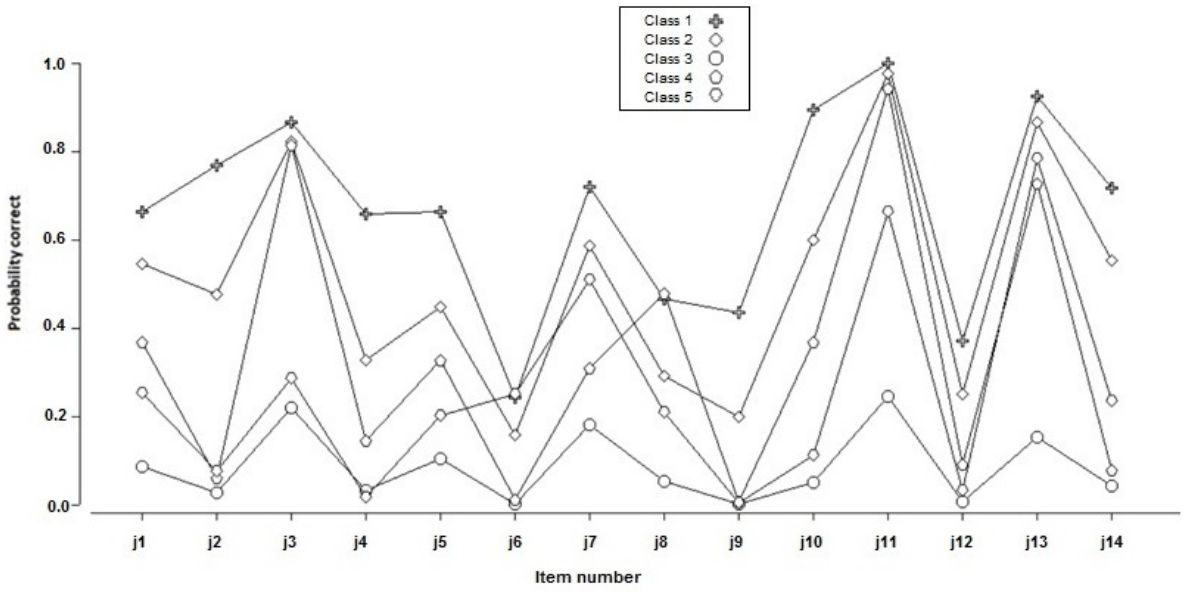
Table 1. BIC Values for One-Class to Five-Class Models Fit to the 2005 Test Data

LCA Solution	BIC	Degrees of Freedom (df)	Classification Errors
One class	213171.81	5286	0.0000
Two class	183986.86	5253	0.0201
Three class	179148.47	5220	0.0537
Four class	177943.63	5187	0.0961
Five class	177994.03	5154	0.1038

In this particular application of LCA, clear interpretation of students' class membership was of high importance, as this was a variable of interest. Therefore, the ability of these solutions to distinguish between classes over the items provided the opportunity to explain membership of Class 1 as generally more proficient in all topics, and the highest class as the least proficient.

For instance, Figure 2 shows that the four- and five-class solutions had many crossings, which made interpretation of the classes difficult. Figure 2 shows 14 items of Study Year Group 1 on the *x*-axis. The mean probability of items being answered correctly for students belonging to one of the five classes is shown on the *y*-axis.

Figure 2. The Profile Plot of 14 Items for the Five-Class LCA Solution

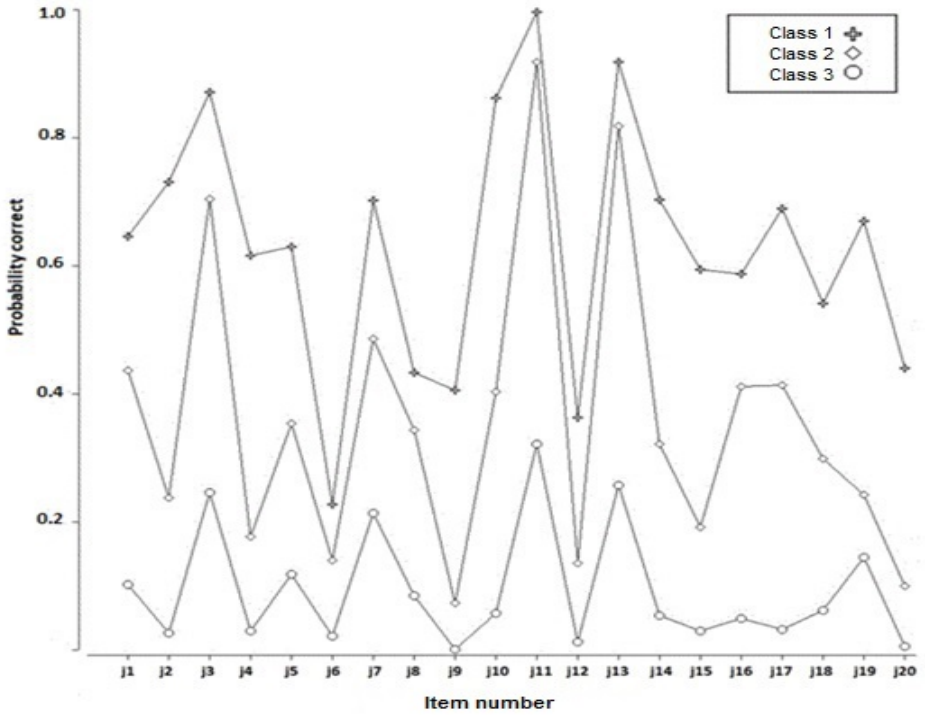


The three-class solution in Figure 3 was better able to distinguish between classes, and the mean probabilities of answering correctly did not cross over items. The same held for the two-class solution. These two LCA solutions were preferred for interpretability of the final outcomes of the class membership. Therefore, both the latent two- and three-class analysis solutions for the different study years (2005–2011) were used in the calibration and simulations for comparison.

The proportion of students per class in the three-class solution can be used to calculate $\bar{P}_c$, which in turn can be used to estimate the restricted group LCAs. Latent class models were fitted separately per year, as the design was not linked. Table 2 shows the proportion of students in the three-class solutions for the different years of the tests.

With the three values of $\bar{P}_{c,}$ for each Latent Class 1 through 3 shown in Table 2, the seven LCA models were again fitted with the restrictions on proportions of students per class. These models give results as conditional probabilities of answering items correctly, which can then be used as item information for an item bank with the three-class solution. Two final item banks were constructed — one for the two-class solution and one for the three-class solution — each

Figure 3. The Profile Plot of 20 Items for the Three-Class LCA Solution



with restrictions on group size. These item banks containing item and person information were used in the subsequent simulation study of adaptive test construction.

Simulations of the Kullback–Leibler selection algorithm for CAPT. In the analysis, 150 answer patterns $\boldsymbol{x_i}$ per class C were simulated for all J items, conditional upon the students' assumed class membership. Subsequently, an item selection algorithm constructed a test of length *t*. Using $\boldsymbol{x_i}$, the class membership could be estimated again, although this time the class

Table 2. Proportion of Students Per Latent Class in the Three-Class LCA Solution

Group/Year	Class 1	Class 2	Class 3
2005	0.377	0.359	0.265
2006	0.316	0.406	0.278
2007	0.363	0.368	0.269
2008	0.351	0.353	0.296
2009	0.388	0.386	0.226
2010	0.345	0.368	0.288
2011	0.385	0.361	0.255
$\bar{P}_c$	0.361	0.372	0.268

prediction was based only on the *t* selected items out of the $\boldsymbol{J}$ items. This procedure was replicated 50 times per class and per selection method. To illustrate, for a two-class solution, $\mathbf{2} \times \mathbf{150} \times \mathbf{50} = \mathbf{15,000}$ answer patterns were generated containing answers to all 208 items in the bank.

Random and KL item selection were compared in terms of proportions of correctly classified students. To evaluate the possible effect of test length on the efficiency of the item selection method, the test length was varied with 10, 15, and 20 items for both methods. The effect of the two different selection methods on the efficiency of classification was of particular interest not only for the two-class solution but also for the three-class LCA model. Note that in both solutions, Class 1 contained the students with the highest probability of answering the items correctly.

The results of the simulation studies are shown in Table 3 in terms of the mean and standard deviation of the number of correctly classified students over the 50 replications for test lengths $t = 10, 15, \text{and } 20$. It can be seen that for both the two- and three-class solutions, the KL selection of 10 items performed better with respect to correct reclassification as compared with random item selection. The two-class solution demonstrated a high proportion of correct classifications when KL information selection was used ($M_{class1} = 0.988, M_{class2} = 0.999$), while the random selection did not perform as well ($M_{class1} = 0.940, M_{class2} = 0.939$). In addition, the standard deviations in the KL information selection were smaller in the two-class solution than in the three-class solution.

Further, for the three-class solution, the KL information item selection method performed better than the random selection method in terms of correctly reclassifying respondents. Class 2 was the most difficult class to predict for both selection methods. Class 1, which contained the best students, had the most correctly reclassified respondents.

Table 3 also shows the results of the simulation with 15 items selected per simulated adaptive test. For both the random and KL information selection methods, the proportion of correctly classified respondents increased in comparison to the 10-item simulated adaptive test. Additionally, the standard deviations were smaller for all cells. For the three-class solution, the differences between the two selection methods decreased.

Finally, for tests of 20 items the mean of the proportion correctly classified over the simulations increased, while the standard deviations decreased. When more items were administered, the random selection method reclassified examinees well ($M_{class1} = 0.986, M_{class2} = 0.987$). However, the KL information showed nearly perfect classification ($M_{class1} = M_{class2} = 0.999$).

Table 3. Mean and Standard Deviation (SD) of Correctly Classified Students after 10, 15, and 20 Items for Two- and Three-Class Solutions

Number of Items	Random		KL	
Solution, and Class	Mean	SD	Mean	SD
10 Items: Two-Class Solution				
Class 1	0.940	0.0212	0.988	0.0098
Class 2	0.939	0.0195	0.999	0.0078
10 Items: Three-Class Solution				
Class 1	0.867	0.0249	0.905	0.0272
Class 2	0.753	0.0323	0.819	0.0348
Class 3	0.863	0.0256	0.873	0.0344
15 Items: Two-Class Solution				
Class 1	0.971	0.0128	0.997	0.0047
Class 2	0.974	0.0117	0.996	0.0056
15 Items: Three-Class Solution				
Class 1	0.918	0.0213	0.948	0.0202
Class 2	0.848	0.0284	0.896	0.0226
Class 3	0.912	0.0221	0.922	0.0183
20 Items: Two-Class Solution				
Class 1	0.986	0.0095	0.999	0.0018
Class 2	0.987	0.0093	0.999	0.0028
20 Items: Three-Class Solution				
Class 1	0.947	0.0186	0.964	0.0128
Class 2	0.905	0.0251	0.933	0.0231
Class 3	0.941	0.0177	0.955	0.0165

The results of the above experiment represent a successful test of the item selection method. It showed that KL information was generally better in estimating class membership than random selection of items. The KL selection method outperformed the random selection method in both the two-class and three-class solution models, and the proportions of correctly specified students increased with test length.

Scoring Based on Posterior Class Probabilities

The difference between IRT and latent class models is primarily in the character of the latent outcome variable in the models. In latent class models, the unobserved outcome variable (ability indication) is categorical. An important question that can arise in practice, however, is

how to use this latent class membership and the corresponding probabilities to provide information on the student's ability. As indicated, three approaches are (1) to use just the class membership, (2) to use the classification in addition to a latent trait ability estimate, or (3) to calculate a score on a continuous scale from the LCA membership probabilities.

Obtaining scores from the classification using a CAT that is based on latent class models comes with some difficulty. The issue that must be solved is how to score students based on their latent class membership. If the goal of the CAT is only to categorize into latent classes, the probability of being a member of a latent class can be used as substitute for a score. If the goal of the test is an outcome on a continuous score, the categorical outcome of the LCA will need to be transformed to a continuous scale.

An interesting property of the latent class model is the probability distribution. Modal class assignment is used to classify students using the probabilities of belonging to the class, given the answer pattern. These probabilities of belonging to class C can also be used to calculate a score on a continuous scale. The new method proposed here for the two-class solution is to calculate the log-odds for a student of the probability of belonging to Class 1 over Class 2, as follows:

$$\text{LogOdds} = \frac{\log(\text{P}(\text{C} = 1|\mathbf{x}_i)}{\log(\text{P}(\text{C} = 2|\mathbf{x}_i)}. \tag{11}$$

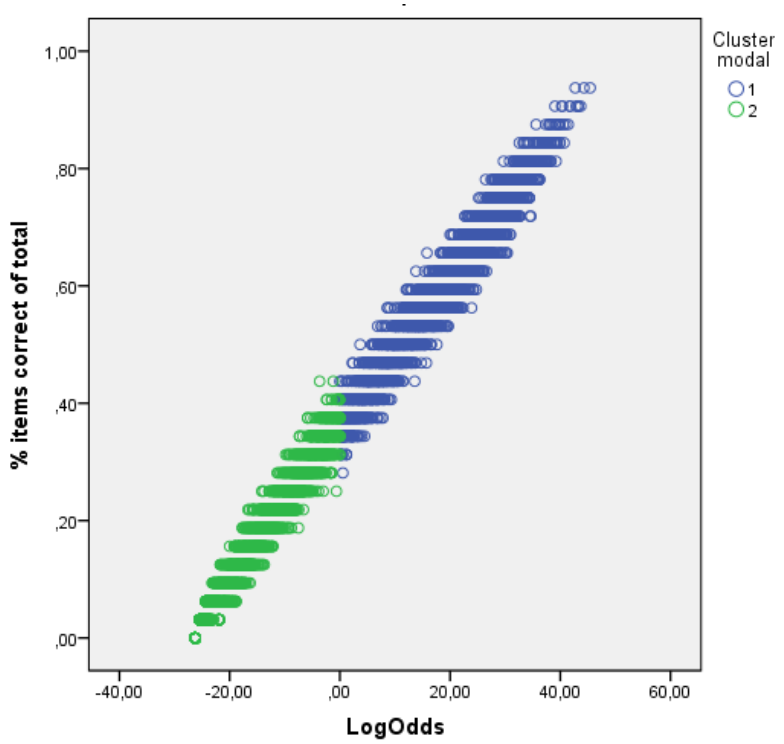
This notation of log-odds would be valid if all students answered the same questions. In an adaptive test, however, the tests are tailored and they differ among students. Thus, the x_i will have to be extended by the item index *k*, which is tailored for each student based on the item selection. To illustrate the purpose of this log-odds scoring, the actual students' scores for the full set of items are shown in Figure 4.

Students with higher and lower levels of knowledge in medical subjects can be recognized by higher and lower log-odds, as the log-odds increases with the percentage of correct items. This increase and the ability to show a relative standing in comparison to other students on a latent continuum is made possible by the log-odds of the two probabilities.

Discussion and Conclusion

This paper has presented item bank construction applying LCA and possible selection and scoring algorithms for CAT. An item selection algorithm based on KL information for the various class distributions has been discussed and applied to an example data set drawn from the Dutch Medical Progress Test. A log-odds scoring procedure based on estimates of class

Figure 4. Log-Odds Scoring for Students in 2015 Contrasted with Percent of Items Answered Correctly



probabilities for this type of CAT has been introduced for a two-class solution. This framework provides the potential to develop CATs using items that do not exactly fit into IRT model frameworks.

The simulation study, in which the KL selection method was compared to random item selection, showed that in all test situations the KL method was more likely to classify students correctly than the random method. KL selection outperformed random selection for all test lengths and for both the two- and three-class solutions. This confirms the earlier study of Cheng (2009), in which KL information was successfully used to optimize cognitive diagnosis processes. A remarkable finding in the simulations is that the random selection method also correctly classified high proportions of students.

The difference between KL and random selection can be seen more clearly in test situations with a smaller number of items than in situations with a larger amount of items. A possible reason for this increase at a longer test length using random selection is that when more items are randomly selected, the probability of selecting an item that contributes well to the classification problem also increases. In addition, the number of classes was still low (two- and three-class solutions) in this application. It is hypothesized that for latent class solutions with a higher number of classes, item selection might benefit even more using KL information selection.

The option of using LCA and efficient item selection methods for CAT applications has been explained in this paper. The outcome of using LCA to deliver CATs will be latent class

memberships, either just a class based on modal classification or an actual probability of belonging to all of the classes. In practice, latent-class-based CAT outcomes parallel to standard CAT measures could be used, which would improve the θ estimates from CAT with a classification. In testing situations where it is not necessary to have a continuous score outcome, the classification could simply be used. An interesting feature is the possibility of working with the probabilities of belonging to a certain class when restrictions have been applied to the LCA. For example, these probabilities can be transformed to actual scores and used in practice.

Further Research

It would be interesting to study more applications of CATs in practice, applying the approach described above. There are many other potential ways to apply the methods presented in this paper. The methods that have been described are capable of being extended to more diverse situations, for example, to developmental psychology.

In addition, improvements can be made to these methods. For instance, the decision on which restrictions to use for the latent class models can be informed using a pseudo-likelihood algorithm to select the optimal class size in LCA. Another very important option for extending this research would be to develop scoring methods for a three-class or higher number class LCA solution, based on the three different conditional probabilities of belonging to one of the classes, given a certain answer pattern. An interesting property of the KL information selection algorithm is that KL information is capable of being extended to a PWKL information criterion (Cheng, 2009, p. 623). The PWKL incorporates prior knowledge about the class to which a student belongs:

$$PWKL_j\left(a_i^{(t)}\right) = \sum_{c=1}^{Nc} D_j(a_i^{(t)} \| a_c)\pi_{i,t}(a_c). \tag{12}$$

There are numerous possible applications of PWKL information, as it would then be possible to use students' previous scores as prior information about their previous class membership for the next progress test a few months later. Using this information measure in an algorithm in future CAPTs would be worth investigating.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123-140. CrossRef

Béguin, A. A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541-561. CrossRef

Chen, S. Y., & Ankenman, R. D. (2004). Effects of practical constraints on item selection rules

at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, *41*(2), 149-174. CrossRef

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619–632. CrossRef

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York, NY: Wiley. CrossRef

Dayton, C. M. (1998). *Latent class scaling analysis*. Sage University Papers Series on Quantitative Applications in the Social Sciences (Book 126). Thousand Oaks, CA: Sage. CrossRef

Eggen, T. J. H. M. (2008). Adaptive testing and item banking. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 215- 234). Cambridge, MA: Hogrefe and Huber Publishers.

Eggen, T. J. H. M. (2012). Computerized adaptive testing item selection in computerized adaptive learning systems. In T. J. H. M. Eggen & B.P. Veldkamp (Eds). *Psychometrics in Practice at RCEC.* Enschede, The Netherlands: RCEC. CrossRef

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates. In J. Hagenaars & A. McCutcheon (Eds.). (2002). *Applied latent class analysis models.* New York, NY: Cambridge University Press. CrossRef

Haughton, D., Legrand, P., & Woolford, S. (2009). Review of three latent class cluster analysis packages: Latent GOLD, poLCA and MCLUST. *The American Statistician*, *63*(1), 81–91. CrossRef

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. CrossRef

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Macready, G. B., & Dayton, C. M. (1992). The application of latent class models in adaptive testing. *Psychometrika*, *57*(1), 71–88. CrossRef

McCutcheon, A. L. (1987). *Latent class analysis*. Beverly Hills, CA: Sage Publications. CrossRef

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with Kullback–Leibler information item selection. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 77–101). New York: Springer. CrossRef

Nylund, K. L., Asparouhov, T., & Muthén, B. (2008). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*(4), 535–569. CrossRef

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research (Expanded edition, 1980. Chicago, IL: University of Chicago Press).

Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: A promising approach for the validation of patient reported outcomes. *Quality of Life Research*, *21*(4), 637–50. CrossRef

Schuwirth, L., Bosman, G., Henning, R. H., Rinkel, R., & Wenink, A. C. G. (2010). Collaboration on progress testing in medical schools in the Netherlands. *Medical Teacher*, *32*(6), 476–479. CrossRef

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6(2)*, 461 464. CrossRef

Suárez-Falcón, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *The British Journal of Mathematical and Statistical Psychology*, *56*(1), 127–143. CrossRef

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*(1). Retrieved from http://pareonline.net/getvn.asp?v=16&n=1

Tio, R. A., Schutte, B., Meiboom, A. A., Greidanus, J., Dubois, E. A., Bremers, A. J. A., & the Dutch Working Group of the Interuniversity Progress Test of Medicine. (2016). The progress test of medicine: The Dutch experience. *Perspectives on Medical Education*, *5*(1), 51–55. CrossRef

Van der Linden, W. J., & Glas, C. A. (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers. CrossRef

Van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*(3), 273–291. CrossRef

Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD user's manual*. Boston, MA: Statistical Innovations.

Vermunt, J. K. (2016). Latent class scaling models longitudinal and multilevel data sets. In G. R. Hancock & G. B. Macready (Eds.), *Advances in latent class analysis: A Festschrift in honor of C. Mitchell Dayton*. Charlotte, NC: Information Age Publishing, Inc.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, C., Chang, H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *37*(2), 1–24. CrossRef

Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: a weighted item selection approach. *Behavior Research Methods*, *44*(1), 95–109. CrossRef

Wrigley, W., Van der Vleuten, C., Freeman, A., & Muijtens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues. *Medical Teacher*, *34*(9), 683–697. CrossRef

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. CrossRef

Xu, X., Chang, H., & Douglas, J. (2003, April). *Computerized adaptive testing strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Quebec.

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177. CrossRef

Author Addresses

Nikky van Buuren, Cito, Amsterdamseweg 13, 6814 CM Arnhem, The Netherlands. Email: Nikky.vanBuuren@cito.nl; Theo H. J. M. Eggen, Cito, Amsterdamseweg 13, 6814 CM Arnhem, The Netherlands. Email: Theo.Eggen@cito.nl

Supplementary Materials

An .Rdata data file called *"LCAsolutions.Rdata"* is available on request from the authors. This file contains:

- An object called "*class2LCA*"; this is a dataframe with the two-class LCA solution output.
- An object called "*class3LCA*"; this is a dataframe with the three-class LCA solution output.