Latent Class Based Algorithms for Computerized Adaptive Testing:
Improving the progress test for medical students in the Netherlands

by
Nikky van Buuren

A thesis defended on July 5th, 2013

at the Mathematical Institute of Leiden University

MSc Mathematics

Statistical Science for the Life and Behavioural Sciences

Advisors:

Prof. Dr. Theo J.H.M. Eggen
CITO in Arnhem

Prof. Dr. Willem J. Heiser
Mathematical Institute, Leiden University

Abstract

In this thesis already existing methods for the construction of a computerized adaptive test(CAT) are extended, with as a goal to create an adaptive test which measures progress. For this project, data is available from seven different medical progress tests by university students in the Netherlands. A CAT, usually has an item response theory(IRT) model applied to the data which can be used to construct an item bank to select items from. However, many items in the medical progress data do not follow an IRT model. Therefore, a new method is proposed to measure progress in a CAT, which is based on a latent class model with 2 or 3 latent classes. The estimated probabilities to answer correctly when belonging to any of these classes are used to calculate Kullback Leibler(KL)-information for all the items. The KL-information values can then be used to select items and to construct an adaptive test. Simulations show that item-selection based on KL-information outperforms random selection of items in progress testing. Finally, a scoring method based on latent class probabilities in the CAT is proposed.

Table of Contents

Introduction 1

Methods 6

2.1 Data structure 6

2.2 Detecting Item Types 7

2.3 Computerized Adaptive Testing and Item Response Theory 9

2.3 Latent Class Analysis 12

2.4 Latent Class Analysis with Restrictions 14

2.5 Kullback-Leibler Algorithm 16

2.6 Simulations of Computerized Adaptive Progress Test with KL selection 18

Results 21

3.1 Data description 21

3.2 Latent Class Analyses with and without restrictions 21

3.2 Kullback Leibler Selection Algorithm for CAPT - Simulations 28

3.3 Scoring Based on Posterior Class Probabilities 32

Discussion 35

Appendix 42

Introduction

Computerized Adaptive Testing(CAT) is a method of testing in which the subsequent question of a test would be the question with the maximum Fisher information of the calculated latent ability estimate. This means each examinee will have a tailored version of a test drawn from a certain pool of questions. The core elements of a computerized adaptive test are the starting criterion for the test, a calibrated item bank, a selection algorithm and the stopping criterion(Wainer, 2000). The development and application of CAT has increased in the past few decades and over time several educational institutions and testing centers have integrated the adaptive methods in their tests(Weiss & Kingsbury, 1984; Van der Linden & Veldkamp, 2004). CAT has many advantages for both examinees and evaluators. These advantages are mainly expressed in growth in efficiency and a better personal experience about the test due to the tailored questions to one's ability level(Eggen, 2008). The goal in this thesis is to apply and adapt currently existing methods and models for CAT, in situations where progress is repeatedly tested while the ability of the examinee is changing.

The medical faculties of five universities in the Netherlands measure the progress of students four times a year. Transforming the current paper-and-pencil test into a computerized adaptive testing is currently investigated by CITO - the institute for educational measurement in Arnhem - and the university medical centers of Maastricht, Leiden, Nijmegen, Groningen, and the VU Amsterdam(AdAPT, 2008). Students in study year 1 through 6 answer the same 200 questions. This testing method causes a large discrepancy in probabilities to answer questions correct per study year, which is due to the diversity of knowledge about the medical subjects in the student population. Ideally the current progress test would be changed into a digital and adaptive version of the progress test; a Computerized Adaptive Progress Test(CAPT).

Frequently used methods to construct a CAT usually first fit a model from item response theory(IRT) to the data in order to construct a calibrated item bank(Embretson & Reise, 2000). With the estimated parameters from these IRT models, tailored item selection can take place for each examinee. Two examples of applying IRT models to a set of items, would be to fit a one-parameter logistic(1-PL) model, which has first been proposed by Rasch(1960), or a two parameter logistic(2-PL) model(Wainer, 2000). These models will be explained more thoroughly in the following sections.

After an item bank is calibrated, an item selection algorithm is applied to continue the construction of the adaptive test. A strategy which can be applied to select the next item in a test, given an estimate of ability $\hat{\theta_i}$ based on previous reponses of examinee i, is the method that selects the item providing maximum information(Van der Linden & Glas, 2000). This method chooses an item $j*$, with $j = 1, ..., J$, which maximizes the item information at the estimated ability level $\hat{\theta_i}$ of the examinee as follows:

$$I_j(\hat{\theta}_i) = \frac{[\mathrm{P}'_j(\hat{\theta}_i)]^2}{\{\mathrm{P}_{j'}(\hat{\theta}_i)[1 - \mathrm{P}_{j'}(\hat{\theta}_i)]\}.} \tag{1}$$

In this formula, $\mathrm{P}_{j'}(\hat{\theta}_i)$ is the probability of a correct response to item j from the calibrated item bank for an examinee i with an current estimate of ability $\hat{\theta}_i$. Each computerized adaptive test also needs a stopping rule. The adaptive test could, for example, be terminated after a fixed number of items, or when a certain precision level has been attained(Wang, Chang, & Boughton, 2012). The decision on a stopping rule depends on many factors, and can result in a fixed-length or variable-length test.

The 1-PL, 2-PL and other extensions of the IRT models all have assumptions and indications of model fit. Not all sets of items adhere to these assumptions and therefore, an IRT model does not always fit. Such an issue is encountered when constructing the adaptive version for the medical progress test. Not all items in the medical progress data fit into one of IRT models. There is a high diversity of knowledge levels among the students, because first-year through sixth-year students all answer the same items. While some items do show a gradual increase in the probability correct over the years, for other items the probability correct among the students demonstrates a sudden jump in a certain year of the study. It is hypothesized that the sudden jump in probability correct in a certain year can be caused by the moment of education of the topic in the curriculum. After this jump has taken place, it is assumed that knowledge of the topic stays constant or only grows slightly over time. Therefore, the main topic of this thesis is exploring the possibilities of constructing a CAPT when having data that contains items in which a sudden jump in the probability correct appears over a period of time. The possible new alternative of using Latent Class Analysis(LCA) in a CAPT construction is investigated in this thesis. Items that do adhere to the normal IRT model are excluded from this part of the CAPT construction. The main research question in this thesis will be: To what extent can Latent Class Analysis be implemented in an attempt to construct a Computerized Adaptive Progress Test for items not following an IRT model?

Latent Class based adaptive tests have never been applied before in progress testing and therefore, all the necessary steps of CAT construction will be discussed. Starting with the construction of a calibrated item bank with item and person information based on latent class parameters, and then finding a selection algorithm and a stopping criterion with adequate achievement. After latent class models are fit to contstruct a calibrated item bank, the next step would be to find a compatible item selection algorithm. The option of using the Kullback Leibler(KL) algorithm (Cheng, 2009) for the items in a latent class framework will be elaborated upon. The selection algorithms used in this thesis are based on the general Kullback Leibler(KL) information measure: $D[f,g] = E_f\left[log\frac{f(\mathbf{x})}{g(\mathbf{x})}\right]$. Cheng and others(Jiao, Macready, Liu, & Cho, 2012) recognized the potential of item selection based on KL-information in CAT for cognitive diagnosis models, which are also based on predictions of attribute profiles which compare to the latent classes in this analysis. The important difference between the study of Cheng, in which cognitive diagnoses are established with adaptive testing, and this study is that the main goal of these progress tests is to give a precise and standardized scoring method. The current progress test wants to score students relative to the cohort they belong to. A method to score persons with help of their latent class probabilities is also proposed, which is an addition to the current studies on CAT with item response ability ($\hat{\theta}$) estimates.

This thesis has been build up as follows. Firstly, the structure of the data and the issue of the sudden jump in the probability correct for items will be explained. Then, latent class analyses with differing amount of classes are compared. Optimal choices about the amount of classes are weighted against each other with several fit indices. After which the outcomes of the latent class models are used in a computerized adaptive test selection algorithm. In the final part of this thesis, the new method to construct a Computerized Adaptive Progress Test(CAPT) with the Kullback-Leibler item selection will be evaluated. This validation is performed by simulation studies in which latent classes are attempted to be estimated from simulated item responses to the progress test. The simulations compare CAPT's with random selection of items to CAPT's with Kullback-Leibler selection of items. Several factors, such as test length and number of latent classes, are varied to compare achievement of both selection algorithms in different test construction situations.

Methods

2.1 Data structure

The data structure of the medical progress test shows item responses of all medical students of the five universities from 2005 through 2011. Since the test is conducted quarterly, one of the four moments of each year is chosen, in this case the December moment. Each of the tests consisted out of 200 items, meaning that a total of 1400 items are available for possible selection in the item bank for the CAPT. The design is disconnected, so there are no overlap or anchoring items which return in the tests over years. Therefore, one can not measure the relative difficulty of the items in the tests over the years. In order to construct an item bank with one underlying scale and to find a calibrated item bank, it is essential to set an assumption about the student population. No identification of students in the data is available, which means longitudinal modelling is not possible. Therefore, an assumption is made that the student populations of all December moments come from the same ability distribution. Hereby, new respondents would be substitute for other respondents leaving the sample.

In the current test, scoring on these 1400 items are either -1, 0 or 1, respectively representing a penalty for a wrong answer, don't know the answer and a correct answer. This penalty for a wrong answer will be removed in the CAPT as this is often seen as a bias to the reliability of the test score(Baldiga, 2012). In the experiment of Baldiga it showed that inserting a penalty option in a test would cause women to answer less questions, while they do have the same knowledge of the material as others. This would give them a disadvantage in the test. However, a study within the AdaPT project showed that changing the -1 penalty into 0, is the best way to handle the data which has been collected under these specific conditions. In this way, the extent to which willingness to take risk is incorporated in the measurements is minimized. This also corresponds to the scoring method that would be used in the computerized adaptive test, since the penalty option will not be incorporated in the CAT.

2.2 Detecting Item Types

In Table 1 one can see amount of items which fall into the different categories that can be distinguished in the data, so called 'grow-type', 'jump-type' and 'rest-type' items. An explanation about how these types are detected can be found below.

Table 1: Classification of Items in Medical Progress Test

Year	**# Grow Items**	**# Jump Items**	**# Rest Items**	**Total**	**# Students**
2005	**56**	**32**	**136**	**200**	**5318**
2006	**38**	**28**	**134**	**200**	**6659**
2007	**41**	**20**	**139**	**200**	**6898**
2008	**48**	**27**	**125**	**200**	**7001**
2009	**27**	**39**	**134**	**200**	**7062**
2010	**51**	**31**	**118**	**200**	**6246**
2011	**41**	**31**	**128**	**200**	**6103**
Total	**302**	**208**	**890**	**1400**	**45287**

Before beginning to construct a model for the items that show a sudden jump it is important to give a clear definition on how this jump in probability correct is detected. An item would be classified as a jump item if the following requirements, with $x \in \{0, 1\}$ and j indicating items, were met:

- P$(x_j = 1|x_j = answered) > .25$ in at least one of the year groups
- P$(x_j = 1|f_j = 6) > .4$, with f = study year, where $f \in \{1, 2, ..., 6\}$
- P$(x_j = 1, f_j = f_{after}) - P(x = 1, f = f_{before}) > .3$, with f_{after} being the year after the subject is teached and f_{before} the year before.

Thereby, these rules show that these items are automatically also accepted when $P(x_j = 1|f = 6) > 0.4$, in correspondence with the first condition specified above. One could imagine that these type of items cannot only occur in the medical test, but might as well occur in other progress tests or tests designed to measure other developments over time. Some groups in the test population are already able to perform well on certain items, while others do not. This exact moment of change in knowledge or moment of development is difficult to indicate, because not for all classes and students the moment of education is the same. Since this thesis only deals with the special case of the jump-type items, one could see in table 1 that 208 jump items are identified over the seven years and will provide the data for the analysis.

Grow type items are defined as items which had at least 10% or 20% increase in probability correct, comparing the percentage correct at the moment of education and at the final - 6^{th} - study year. It compared the probability correct between the subject was teached in year 1 through 3(at least 20%) or year 4 through 6(at least 10%)

2.3 Computerized Adaptive Testing and Item Response Theory

Fitting item response theory(IRT) models to a set of items is a common way to construct a calibrated item bank which can be used in a computerized adaptive test(CAT). In order to give a thorough understanding of the topic, the current methods of using IRT models to construct a CAT are outlined in this section. Section 2.4-2.6 will provide insights on the possible new methods proposed in this thesis in constructing a CAT. The main goal of a CAT is to create an algorithm that tailors the test in a way that it is a valid measure of a particular individual. The invariance property of IRT models makes this possible (Embretson & Reise, 2000).

The general idea of adaptive testing is to avoid inefficiency of methods of testing in which all student have to answer the same items. Examinees with high trait levels will most probably answer many easy items correctly, where this does not give any information on their relative standing. On the other hand, lower trait level examinees will answer difficult items wrong and these also do not contribute much to the scoring on a relative scale. This is the general idea of a computerized test; to offer items tailored to the latent ability level of an examinee.

A unidimensional IRT model is supposed to measure only one latent trait, which in educational testing would be measuring one value of latent ability that should be sufficient to characterize differences between persons. There are several unidimensional models available, but only the one-parameter logistic model(1PL) or Rasch model and the two-parameter logistic model(2PL) will be discussed here.

The Rasch model gives a prediction of the probability of a correct answer of person i to item j:

$$\mathrm{P}(x_j = 1|\theta_i, \beta_j) = \frac{\exp^{(\theta_i - \beta_j)}}{1 + \exp^{(\theta_i - \beta_j)}}. \tag{2}$$

The Rasch model is extended to the two-parameter logistic model by adding an extra parameter, namely the parameter α_j. This value α_j allows for differences in items for the discrimination, and then the model is of the following form:

$$\mathrm{P}(x_j = 1|\theta_i, \beta_j, \alpha_j) = \frac{\exp^{\alpha_j(\theta_i - \beta_j)}}{1 + \exp^{\alpha_j(\theta_i - \beta_j)}}. \tag{3}$$

In the Figure below, an overview of examples of item characteristic curves of IRT models. The left model shows a model in which the discrimination level α_j differs, but the difficulty level β_j remains constant. In the right figure, the difficulty level varies but there is no discrimination level that fluctuates(1PL). In these models, examinees with a certain level of the latent trait θ answering an item with difficulty level β_j have a probability to answer that item correctly of 0.5 if $\theta = \beta_j$.

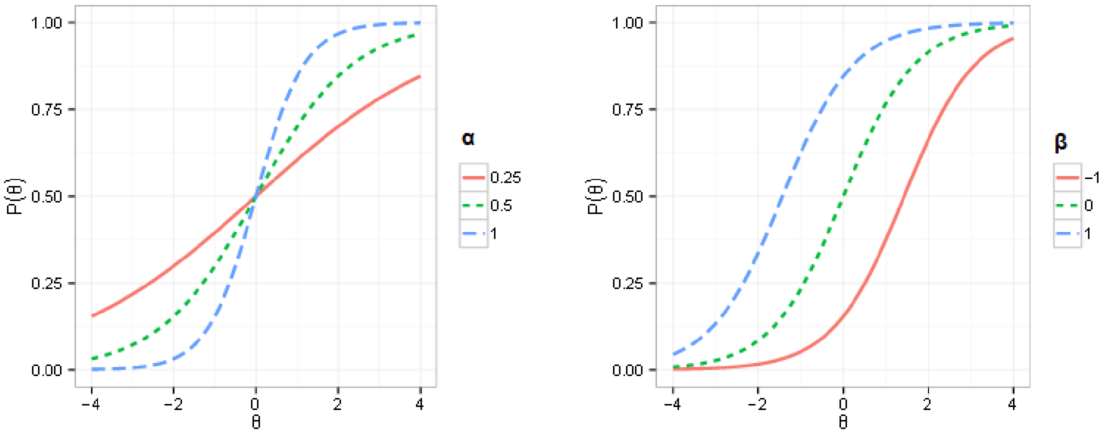

Figure 1: Item Characteristics Curves 1PL and 2PL

When the problem occurs of proposing too easy or too difficult items to examinees, one could make use of a computerized adaptive test which usually attempts to administer a test in which the proportion of correctly answered items is around 0.50(Embretson & Reise, 2000). The likelihood function of the student's ability plays a large role in making inference on the student, since it makes it possible to estimate the ability of the student for each subset of items on the same scale. Given the scores on J items with $j = \{1, ..., J\}$, the likelihood of the answer pattern is as follows:

$$L(\theta; x_1, ..., x_J) = \prod_{j=1}^{J} p_j(\theta)^{x_j}(1 - p_j(\theta)^{1-x_j}) \tag{4}$$

This likelihood function is used to compute estimates, but can also be used for statistical testing. The maximum of the likelihood function(MLE estimate) is often used as the value for ability level θ. The standard error of the estimated θ can also be calculated and is often used as a stopping criterion for a computerized adaptive test. Item selection takes place with help of the Fisher information function [formula 1], which is defined as the statistical expectation of the squared relative change of the likelihood function at the ability level $\hat{\theta}_i$. After each item answer, the item with the maximum Fisher information is the best fitting item for the examinee. When the stopping rule of a pre-set level of the standard error of the ability estimate is used, the CAT is usually stopped when this level is reached. Then, the final estimates can be calculated(Wang, Chang, Boughton, 2012).

2.3 Latent Class Analysis

Latent Class Analysis(LCA) is a method of analysis in which observable or manifest variables are related to unobservable or latent variables. Both the observed and unobserved variables in this model are assumed to be categorical (Lazersfeld & Henry, 1968). This is the essential difference when compared to psychological measurement models as item response theory(IRT), because in IRT the observed variables would also be categorical, but the underlying latent variable would be continuous. Basically, LCA models give us a probabilistic or fuzzy outcome on predicted class membership, while 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 on 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 $\mathrm{P}(C = c|\mathbf{x}_i)$ given the observed answer pattern $\mathbf{x}_i = (x_{i1}, ..., x_{iJ})$ of an examinee i can be calculated, with C being the total number of latent classes(Dayton, 1998). With only dichotomous items, where $x_{ij} \in \{0, 1\}$, and $i = 1, ..., n$ for the persons and

$j = 1, .., J$ for the items. A latent class is indexed by c, with $c = 1, ..., C$. The joint probability of obtaining a certain response pattern x_i would be:

$$\mathrm{P}(\mathbf{x}_i) = \sum_{c=1}^{C} \pi_c \prod_{j=1}^{J} \mathrm{P}(x_{ij} = 1 | C = c). \tag{5}$$

Here, π_c indicates the proportion of students that belong to class c. In formula [5], one can recognize the assumption of local independence. Because the J observed items are assumed to be mutually independent within each latent class c. An interesting aspect for this thesis and its use of LCA is that it assigns students to a latent class with a certain probability. The estimated class probability for each individual to belong to a class C given someone's answer pattern $\mathbf{x}_i$ would be

$$\mathrm{P}(C = c | \mathbf{x}_i) = \frac{\mathrm{P}(\pi_c)\mathrm{P}(\mathbf{x}_i | C = c)}{\mathrm{P}(\mathbf{x}_i)}, \tag{6}$$

which are referred to as posterior probabilities. To illustrate, for a correct answer to an item for a person, this would result into the following expression: $\mathrm{P}(C = c|x_{ij} = 1) = \frac{\mathrm{P}(x_{ij}=1|C=c)\pi_c}{\mathrm{P}(x_{ij}=1)}$. In LCA, each observed response pattern $\mathbf{x}_i$ would belong with a certain probability to each of the classes. In case the goal of the latent class analysis is person classification, the most common used classification rule is modal assignment. This rule means that the examinee will be assigned to the class with the highest $\mathrm{P}(C = c|\mathbf{x}_i)$(Hagenaars& McCutcheon, 2002). Restrictions can be applied to LCA's, which are discussed in section 2.6.

For the LCA analyses in this thesis, the program used is Latent Gold 4.5(Vermunt & Magidson, 2000; Haughton, Legrand & Woolford, 2009). Latent Gold makes use of Maximum Likelihood (ML) or Posterior Mode(PM) estimation methods to estimate the parameters of the latent class models. To find these ML or PM estimates for the model parameters, Latent Gold uses both EM as well as the Newton-Raphson algorithm.

In order to be able to draw conclusions on how many classes are the best fit to the data and what would be a good choice for the class sizes, several indices can be looked at. The first index measure and global fit index that can be looked at is the -2 Log likelihood(-2LL) of the models with different number of cluster(Nylund, Asparouhov, & Muthén, 2008).

Another criterion is the Bayesian Information Criterion (BIC; Schwartz,1978), which in many articles and studies is regarded as a good indicator of distinguishing between models with different numbers of latent classes(McCutcheon, 1987; Vermunt & Magidson, 2000; Hagenaars & McCutcheon, 2002). The BIC is defined as:

$$BIC = -2 \log L + p \log(n)$$

,where p is the number of free model parameters and n represents the sample size. This information criterion also uses the -2 log likelihood, but adjusts for the amount of parameters and the sample size in the model as well resulting in a more reliable index(Nylund et al., 2008).

2.4 Latent Class Analysis with Restrictions

There are multiple latent class models that should be fit to the progress test data, because of the disconnected data structure(section 2.1). However, the main goal is to use latent class analysis to build one calibrated item bank. This would actually mean that all the G seperate models with $g \in \{1, ..., G\}$, corresponding to year 2005 to 2011, have to be connected. This connection of the different models can be made by assuming that the characteristics of the student population do not change over the years. Supported by this assumption, one can fix latent class sizes within each model to a constant class size and fit all the models again with this restriction. By setting this assumption, one can fit seperate LCA models for each year group g, and still use the items of different years in the same item bank. The application of this assumption is outlined below:

Model 1

This model has the following notation; group is indicated by $g \in \{1, ..., G\}$, and an item by $j \in \{1, ..., J_g\}$ with J_g being the total number of items in group g. Latent class is indicated by $c \in \{1, ..., C\}$, with C again being the total number of latent classes. This results in the following model, which is an extension of the non-group specific formula [5]:

$$\mathrm{P}(x_{i1}^{(g)}, ..., x_{nJg}^{(g)}) = \sum_{c=1}^{C} \pi_c^g \prod_{j=1}^{J_g} \mathrm{P}(x_{ij}^g = 1 | C = c), \tag{7}$$

in which a restriction can be applied(Macready & Dayton, 1992). Then, use model 1 in the following algorithm:

```
Algorithm
Step 1
For $g \in 1, ..., G$ , compute maximum likelihood estimates of model parameters
Parameters $= \arg\max_{L_g(model1)}$
Step 2
$P_c^g = \pi_c^g$
$\bar{P}_c = \frac{1}{G} \sum\limits_{g=1}^{G} P_c^g$ yields the average class size over all groups g for all $c \in \{1, ..., C\}$
$P_c^L = \min_g P_c^g$ , yield the minimum class sizes for $c \in \{1, ..., C\}$ of all groups $g$
$P_c^U = \max_g P_c^g$, yield the maximum class sizes for $c \in \{1, ..., C\}$ of all groups $g$
Step 3
For $g = 1, ..., G$, estimate model A, and this time with extra restriction $P_c^g = \bar{P}_c$
```

In step 2 of the algorithm, one calculates the average group size $\bar{P}_c$ that is used to connect the scale. It is, however, also possible to find a pseudo likelihood estimate of P_c^g. This procedure of finding pseudo likelihood estimates is beyond the scope of this thesis.

In step 3 of the algorithm, one estimates the parameters in the models again, but this time with fixed class sizes over all years g based on the computed average group size $\bar{P}_c$.

2.5 Kullback-Leibler Algorithm

To compute the Kullback-Leibler Information (Kullback & Leibler, 1951; Cover & Thomas, 1991) between two probability distributions, the general form is expressed as follows:

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

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 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, since the measure is not symmetric with $D[f,g] \neq D[g,f]$. Large KL-information $D[f,g]$ would be an indication of two statistically different distributions, simply indicating deviances between the two (Cheng, 2009). There are several terms used in literature, i.e. KL-distance, KL-divergence and KL-information. The latter term is the one chosen to be used in this thesis.

An application of the Kullback-Leibler algorithm which helps to select items in order to optimilize cognitive diagnosis has been proposed by Cheng(2009). In Cheng's application, the set of t available items is referred to as $R^{(t)}$, with item j again indicating an item. Interest in cognitive diagnosis is to classify persons into cognitive profiles α_i. The true state of a person's attribute profile is unknown and therefore the KL distance between 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 formula [8], one could calculate the KL information between $f(\mathbf{x}_i|\hat{\alpha}_i^{(t)})$, with $\hat{\alpha}_i^{(t)}$ indicating the current estimate of $\hat{\alpha}_i$ and another distribution based on $\mathbf{x}_i$ given another latent state α_c:

$$D_j(\hat{\alpha}_i^{(t)}||\alpha_c) = \sum_{q=0}^{1} log\left(\frac{\mathrm{P}(\mathbf{x}_{ij} = q|\hat{\alpha}_i^{(t)})}{\mathrm{P}(\mathbf{x}_{ij} = q|\alpha_c)}\right)\mathrm{P}(\mathbf{x}_{ij} = q|\hat{\alpha}_i^{(t)}). \tag{9}$$

This would be the case for dichotomous items. Then, if more than one alternative latent state is available, one could use the sum of the KL information between the estimated current latent state $\hat{\alpha}_i^{(t)}$ and all other latent states a_c's (Xu et al., 2003) :

$$KL_j(\hat{\alpha}_i^{(t)}) = \textstyle\sum_{c=1}^{C} D_j(\hat{\alpha}_i^{(t)}||\alpha_c). \tag{10}$$

Item selection would then follow from the idea of finding the item with the maximum $KL_j(\hat{\alpha}_i^{(t)})$ 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. In case there are two latent states, calculating the KL-information as in the formula [9] already suffices.

Cheng (2009) also proposes extensions of the above described KL-information, of which one is the posterior weighted KL-information(PWKL). The PWKL could incorporate information about old samples in the analysis of current samples with help of prior information. Since the progress test data has no identification of students over the years, the PWKL is for this moment not possible.

2.6 Simulations of Computerized Adaptive Progress Test with KL selection

In this section the set-up for the simulations is outlined for both the adaptive tests with random item selection and with KL-information item selection. To simulate

Computerized Adaptive Progress Tests(CAPT's), response patterns were generated conditioned upon latent class membership. In each simulated CAPT, 150 examinees per latent class where assigned to their 'true' latent state. Basically, this means that the two-class LCA solution simulation tests the achievement of reclassifying 300 examinees based on the simulated response patterns, while the three-class solution tested the classification accuracy for 450 examinees. In the CAPT simulation, item responses were generated based on item response probabilities given the 'true' class that the examinee belongs to. These item response probabilities are already estimated in the restricted latent class models described in section 2.5. Replications of these simulations makes it possible to draw conclusions about variance in the different test situations.

A random generator, from the uniform distribution with $d \sim U(0,1)$, was used to simulate these answer patterns. From this distribution, for each of the 50 replications a total of D uniformly distributed numbers are drawn. The 208 jump items in the databank and the LCA 2-class and 3-class solutions based on these items have the estimated probabilities $\mathrm{P}(x_j = 1|C = c)$. To generate answer patterns, for the uniformly distributed d values, the probabilities $\mathrm{P}(x_j = 1|C = c)$ have been compared to the boundary values d,

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

The simulated answer patterns obtained in previously described manner, are then again used to estimate the latent class membership from. In contrast, this time only a subset of the item responses are used to estimate. The idea of the CAPT is to use less items than the current 200 items progress test and come to a correct prediction of the true latent class and latent ability of the examinees. Two methods of item selection are compared, the baseline would be random selection of the items and the experimental setting which is expected to increase the performance of the CAPT is selection based on the Kullback Leibler(KL) information.

The random selection method generates for each examinee a random vector of test length 10, 15, or 20 out of the 208 jump-type items. The KL method also generates different tests for each examinee, but the items are chosen based upon the current class estimate $\alpha_i^{(t)}$. After each question the current estimate of the class α_i^t is used to choose the best fitting item according to the KL information indices. As described in section 2.5, the predicted latent class probabilities are updated after an item is administered and with the modal classification rule(see section 2.3) the next item is chosen based on the largest KL information corresponding to the predicted latent state $\alpha_i^{(t)}$. In contrast to the CAT with an underlying item response model where the ability parameter on a continuous latent scale is used to select the optimal next item, the CAPT for the jump items will in this case use the categorical class membership estimate to select the next item. The items with the maximum the KL information are supposed to contribute most to the correct classification of the examinees.

Results

3.1 Data description

Most of the students who took the progress test were in their 1^{st} to 3^{rd} year of the medical program(58,7%). Over the years 2005 through 2011 the sample sizes differed between a minimum of $n = 5318$ in 2005 and a maximum of $n = 7001$ students in 2008. Identification of individual students is not possible and missing data did not occur.

3.2 Latent Class Analyses with and without restrictions

Since there is unconnected data from seven different year groups, there are seven latent class models fitted to the data. Firstly, these models have been freely estimated, and different amount of classes have been tested and compared. With help of -2LL, BIC and the misclassification error rates, the best number of latent classes can be compared and be used as a general framework for the simulations of the progress tests in the next section. The results of these freely estimated latent class models are then used to impose restrictions on class sizes. This is necessary to obtain a connected item bank, which can be used in the construction of an adaptive test.

In this section of the thesis, only the results of year 2005 - one of the seven year groups - are elaborated upon. Fit indices that are looked at in order to compare to models with more clusters, is the -2 log likelihood, the BIC and the misclassification errors.

Table 2: Fit indices of different amount of classes in 2005

2005	**LogLikelihood**	**df**	**BIC**	**Class. errors**
1-class	-106448.64	5286	213171.81	0.0000
2-class	-91714.62	5253	183986.86	0.0201
3-class	-89153.87	5220	179148.47	0.0537
4-class	-88409.89	5187	177943.63	0.0961
5-class	-88193.55	5154	177994.03	0.1038

It is shown that the 4-class solution is preferred by the BIC indicator, but as of the 4-class solution large misclassification errors seem to appear. The goal of this thesis is to deliver a reliable and valid classification method for testing. Therefore, random fluctuations in uninterpretable cluster classifications should be avoided. This is the reason why the 2 and 3-class solution are to be preferred based on the good interpretability with respect to content and the lower amount of misclassifications. An example of the problem with more than three classes is shown below in Figure 2, with some obvious crossings within the classes and items. There is no theoretical interpretation for these crossings.

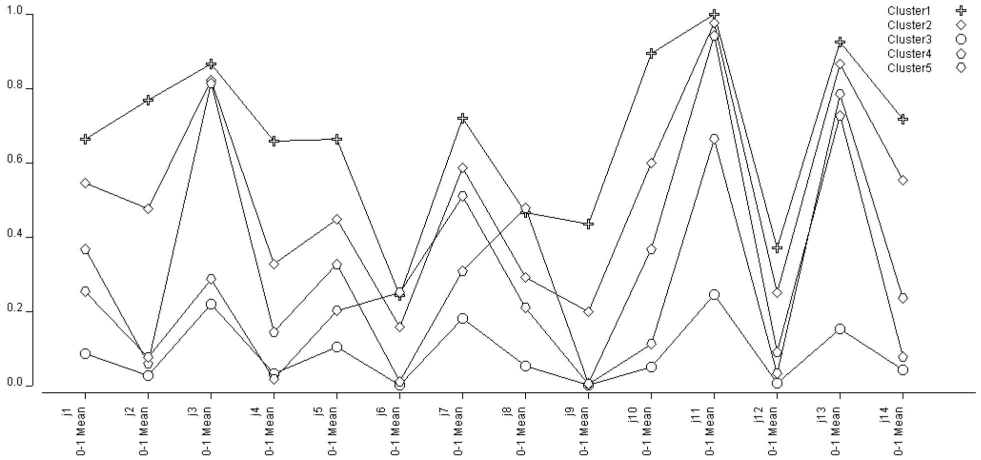

Figure 2: The profile plot of the 5-class solution

2-Class solution

The analysis continues with extending and constructing correct models for these 2 and 3-class solutions. Figure 3 is showing the average probability per item to answer correctly per cluster for the 2-class solution. The clusters are formed based on similarities in the population. The most capable students, with most correct item responses, will end up in cluster 1. The clusters with the higher numbers contain the students with less items correct.

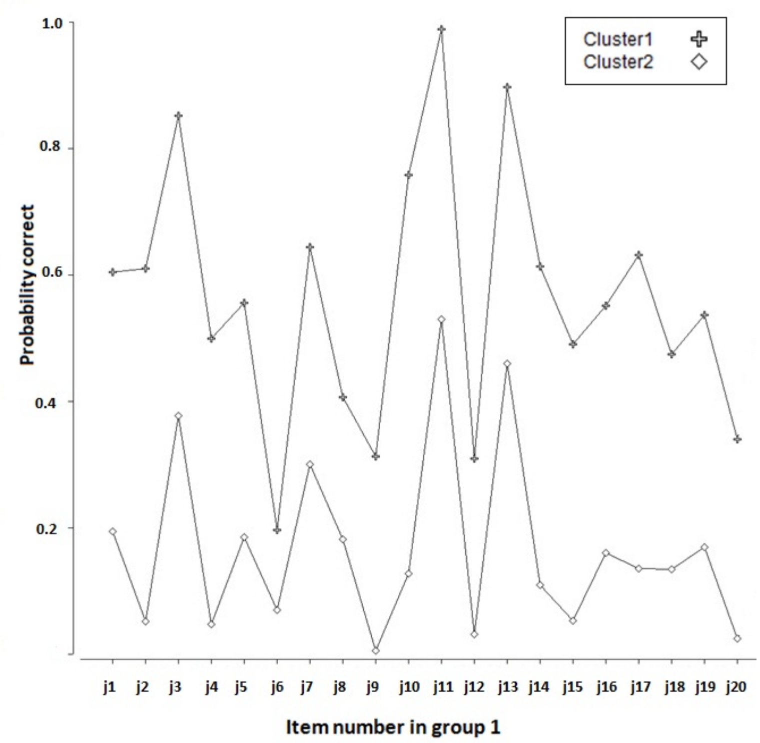

Figure 3: Profile Plot of the 2-class solution of 2005

The profile plot indicates for example a probability of .604 to correctly answer item 1, given the cluster membership of cluster 1. When member of cluster 2, this probability is only .194. Exact probabilities for all items of 2005 can be found in de Appendix [table 1].

The following class sizes are found for the freely estimated latent class model with two classes over all the seven years 2005-2011.

Table 3: Proportion of students per class in the 2-class solution LCA

Class sizes	**Class 1**	**Class 2**
2005	0.567	0.433
2006	0.535	0.465
2007	0.537	0.463
2008	0.529	0.471
2009	0.549	0.451
2010	0.536	0.464
2011	0.570	0.430
$\bar{P}_c$	0.536	0.464

The main finding from these seperately estimated models is that the class sizes remain approximately constant over the years. This would also be a support for the assumption of the population coming from one distribution.

To continue, the algorithm from section 2.4 is applied to find the $\bar{P}_c$ for $c = \{1, 2\}$ over all groups g for the two class solution. With the values of $\bar{P}_c$ for the two class solution, latent class models with a restriction on class sizes are then defined. An additional restriction in these subsequent fitted models is that the students with the highest probability to answer items correctly should always end up in class 1. The values of $\bar{P}_c$ are calculated over all seven year groups(2005-2011) and are fixed in the subsequent analyses by: $\bar{P}_1 = 0.536$ and $\bar{P}_2 = 0.464$. The latent class probabilities coming from these restricted latent class models together form the connected item bank for the 2-class solution.

3-Class solution

The 3-class solution has also been plotted, and is shown in Figure 4 below. For some items, the probability for the lower class 3, seems to be extremely low. In the Appendix[table 2], one can find the exact conditional probabilities. The probability to answer item 20 correctly given membership of latent class 3 , for example, is only .0056. Being member of class 1, gives an estimated probability of 0.4392 for the same item. On the other hand, members of class 3 already have a probability of .3208 to answer item 11 correctly, and in the classes with higher levels this increases to really high probabilities(Class 2: .9179, Class 3: .996). What one can basically see in the results, is a good seperation of knowledge level between three classes.

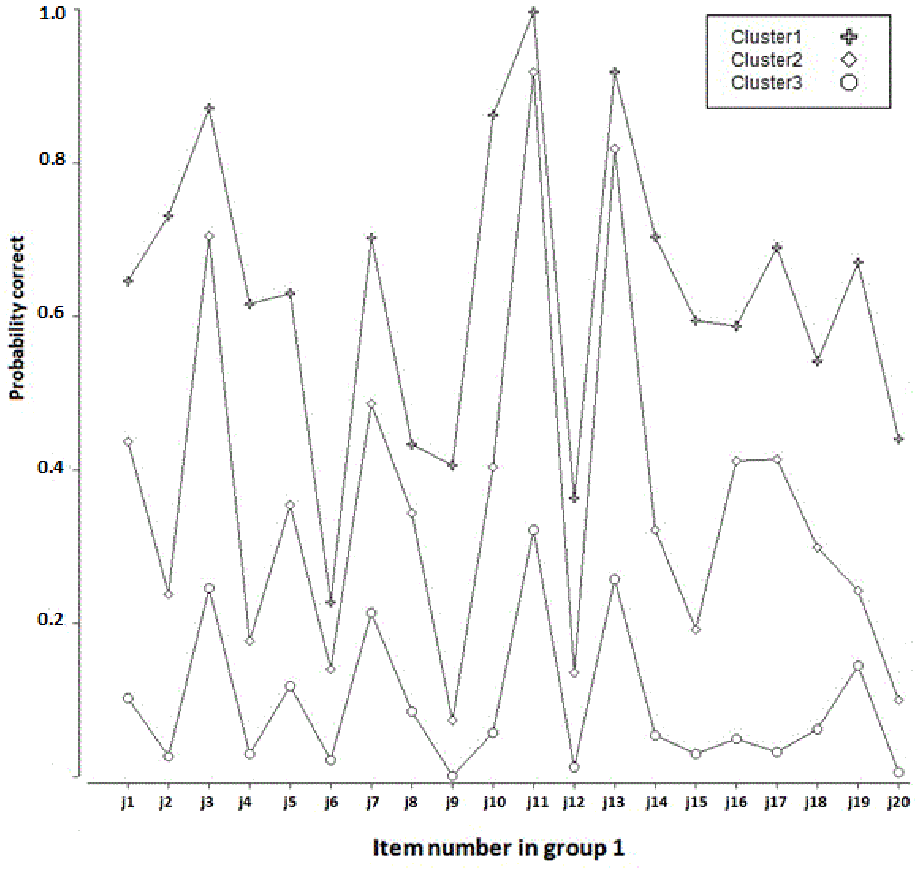

Figure 4: The 3-class solution of the LCA on the first 20 items of 2005

For the three class solution, the average class sizes $\bar{P}_c$ have also been calculated over the years g , as in step 3 of the algorithm in section 2.4.

Table 4: Proportion of students per class in the 3-class solution LCA

Class Sizes	**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

With the values of $\bar{P}_1, \bar{P}_2, \bar{P}_3$, which are shown in the table above, the seven LCA models are again fitted with these restrictions. These models give results as conditional probabilities to answer correctly to items, which are now used as item information for an item bank. Two item banks have finally been constructed, one for the 2-class and one for the 3-class solution. These item banks containing item and person information are used in the subsequent simulation study of constructing adaptive tests.

3.2 Kullback Leibler Selection Algorithm for CAPT - Simulations

In the analysis 150 answer patterns $\mathbf{x}_i$ per class C have been simulated for all J items, conditioned upon their class membership. After this, an item selection algorithm constructs a test of length t and by using $\mathbf{x}_i$ the class membership can be estimated again, but this time the class prediction would only be based on the t selected items out of the J items. This procedure has been replicated 50 times per class and per selection method. To illustrate, for a two class solution, in total $2*150*50 = 15,000$ answer patterns have been generated and one of the item selection methods is used and it's efficiency is measured in terms of correctly classified students.

In the simulation study there are two selection methods which are to be compared, namely the random selection method of items and the Kullback Leibler selection method as proposed in section 2.6. To have an overview of the possible effect of test length on the efficiency of the selection method, test length has been varied with 10, 15 and 20 items. It is also of high interest to look at the effect of the two different selection methods in the efficiency of classification for not only the two-class solution but also for the three class LCA model. Note, that class 1 contains the students with the highest probability to answer items correctly.

The results of the simulation studies are shown below in the tables, which give a mean of the correctly specified simulated students over the 50 replicates and a standard deviation around this mean for $t = 10,\ 15,$ and 20. In Table 5, it is shown that for both the 2 and 3-class solutions, the Kullback-Leibler item selection of 10 items performs better with respect to correctly reclassifying then the random item selection. The 2-class solution already has a high proportion for the KL-information selection (M_{class1} = 0.988, M_{class2} = 0.999), but the random selection has lower percentages correctly classified (M_{class1} = 0.940, M_{class2} = 0.939). Also standard errors are smaller for KL-information selection in the 2-class solution.

Table 5: Correctly classified after 10 items

Selection method		Random	Random	KL	KL
		mean	*s.d.*	*mean*	*s.d.*
2 Class Solution	**Class 1**	0.940	0.0212	0.988	0.0098
	Class 2	0.939	0.0195	0.999	0.0078
3 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

Also for the 3-class solution, the KL-information item selection method is performing better than the random selection method in terms of correctly reclassifying respondents. The middle class seems to be the most difficult class to predict for both selection methods. The first class, which contains the best students, has many correctly reclassified respondents.

Below, one can find the results of the simulation with 15 items selected per simulated adaptive test. Results show, that for both the random and KL-information selection method the proportion correctly classified increased in comparison to the 10 item simulated adaptive test. In addition, the standard deviations also become smaller for all cells. For the three class solution, the differences between the two selection methods decreased.

Table 6: Correctly classified after 15 items

Selection Method		Random	Random	KL	KL
		mean	*s.d.*	*mean*	*s.d.*
2 Class Solution	**Class 1**	0.971	0.0128	0.997	0.0047
	Class 2	0.974	0.0117	0.996	0.0056
3 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

In the following plot one can graphically compare the achievement of the two item selection methods for an adaptive test with 15 items. The KL-information item selection outperforms the random selection for all classes.

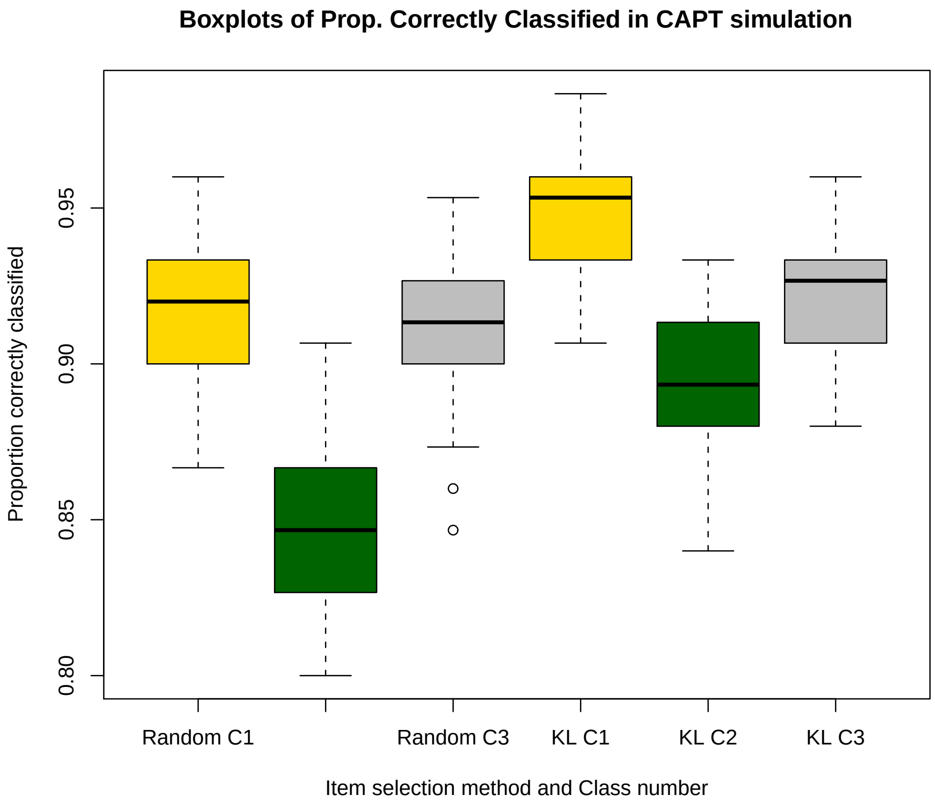

Figure 5: Boxplot of Comparison between item selection methods for test length t=15
The boxplot indicates the distribution of the proportion of correctly classified examinees in the simulation outcomes of a test length of 15 items. The yellow class indicates the first class, green the second class and grey the third class. Subsequently, one can compare the random item selection and KL item selection with each other. For all three classes, the KL selection performs better than the random selection.

Finally, adaptive tests of 20 items are simulated. Again, mean of proportion correctly classified over the simulations has increased and the standard deviations decrease. When more items are administered, also the random selection method seems to reclassify examinees well($M_{class1} = 0.986$, $M_{class2} = 0.987$). However, the KL-information now classifies close to perfect($M_{class1} = M_{class2} = 0.999$).

Table 7: Correctly classified after 20 items

Selection Method		Random	Random	KL	KL
		mean	*s.d.*	*mean*	*s.d.*
2 Class Solution	**Class 1**	0.986	0.0095	0.999	0.0018
	Class 2	0.987	0.0093	0.999	0.0028
3 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 succesful test on the reliability of the item selection method. The Kullback Leibler selection method has outperformed the random selection method. This is the case for both the 2-class and 3-class solution models and the proportions of correctly specified students increase with the test length.

3.3 Scoring Based on Posterior Class Probabilities

In the method section on item response theory(IRT) models, it is emphasized that the difference between IRT and latent class models mainly expresses itself in the character of the latent outcome variable in both models. In latent class models, the unobserved outcome variable(ability indication) is categorical. In section 3.2, the simulations showed that with the KL-information selection method the number of correctly specified students is considerably high. An important issue which arises, however, is to get valid estimates of ability from the classification by the computerized test based on latent class models. In short, the issue to be solved is how to score students based on their latent class membership.

An interesting property of the latent class model is the probability distribution which comes out of the model. Modal class assignment is the way to classify the students, and this is done with help of probabilities of belonging to the class given the answer pattern. These probabilities to belong to class c , can also be used to calculate a score on a continuous scale. The new method proposed here for the 2-class solution is to calculate a log odds for a student of the probability to belong to class 1 over class 2, as follows:

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

This notation of the log odds would be valid if all students would answer the same questions. Since this is an adaptive test, tests are tailored and differ over the students. So the $\mathbf{x}_i$ will have to be extended by the item index k which is tailored for each student based on the item selection. The validity of the item selection method can then be tested by checking if these scorings are actually measuring what they intend to measure.

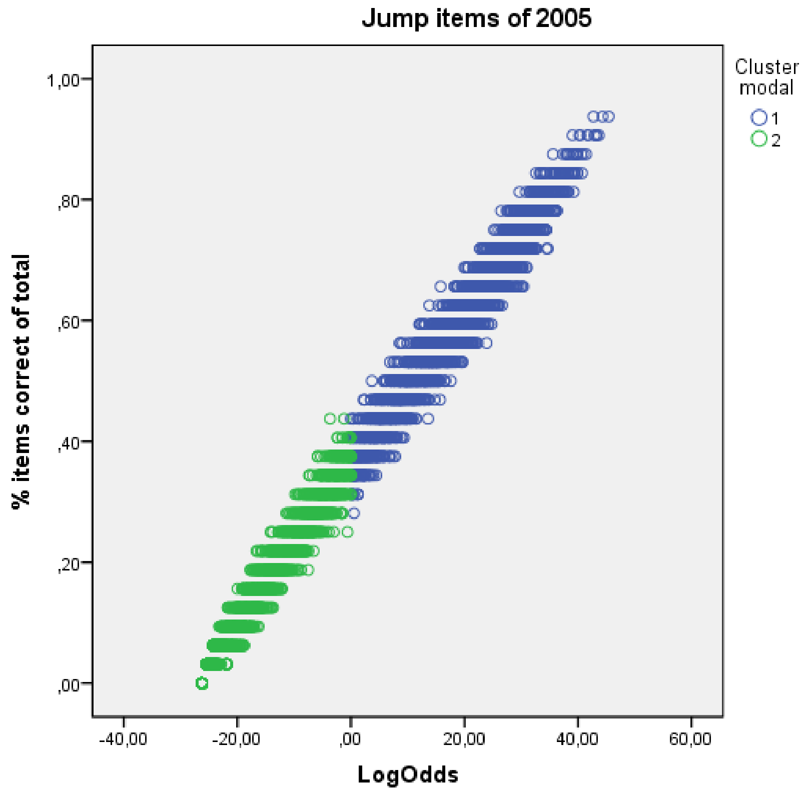

Figure 6: Scoring of 2-class solution plotted against the total score on the items

To illustrate the purpose of this logodds, the student scorings for the full set of jump items have been calculated and these have been plotted against the percentage of items correct to have a global impression on how it follows the scale. Figure 6 shows the respondents classified in cluster 1 and cluster 2, respectively the students with higher and lower levels of knowledge from the medical subjects. Then, the logodds seem to increase with the percentage of items correct. This increase and the possibility to show a relative standing in comparison to other students on a latent continuum is made possible by the logodds of the two probabilities.

Discussion

This thesis consisted out of several comparisons between existing and newly developed methods for item selection in computerized adaptive testing. The possibility of constructing a Computerized Adaptive Progress Test(CAPT) for medical students has been explored, and a simulation study on the possible new item selection method for items showing a sudden jump in probability correct is performed. Additionally, a proposal to transform the categorical classification of the students to a scoring on continuous scale is given. The simulation study in section [3.2], which has compared the Kullback Leibler(KL) selection method with the random selection method, showed that in all test situations the KL more often correctly classifies the students than the random item selection. The KL selection outperformed the random selection for all test lengths and for both the 2 and 3-class solution. This corresponds to earlier findings of Cheng(2009), where the Kullback-Leibler information was successfully used to optimize cognitive diagnosis processes.

Computerized adaptive tests usually have an underlying calibrated item bank with item parameters coming from an item response theory(IRT) model. In the construction of an adaptive test for the university medical program in The Netherlands, a difficulty arose in the data which had to do with non-gradually increasing knowledge levels in the population. These so-called 'jump-items' caused problems in constructing an IRT calibrated item bank. The item selection could therefore not be based on the Maximum Fisher Information criterion. The latent class models seemed to fit the data to a certain extent, but more than 3-class models had high levels of misclassifications. The 2 and 3-class solutions were therefore chosen as the basis for the simulated computerized adaptive progress tests. A solution for the identificationproblem in the non-overlapping data is found by taking average class sizes and assuming students from different years come from the same distribution.

A remarkable finding in the simulation, is that the random selection method of items is also correctly classifying large proportions of students. The difference between KL and random selection can be seen more clearly in the test situation with smaller amount of items, than with larger amount of items. A possible reason for this increase with longer test length, at the random situation is that when items are randomly selected the probability of selecting an item that contributes well to the classification problem also becomes higher.

An interesting property of the Kullback Leibler information selection algorithm is that the KL information could be extended to a Posterior Weighted Kullback Leibler(PWKL) information criterion(Cheng, 2009, p.623). The PWKL incorporates prior knowledge about the class to which a student belongs, as follows;

$$PWKL_j(\overset{\wedge}{\alpha}_i^{(t)}) = \sum_{c=1}^{C} D_j(\overset{\wedge}{\alpha}_i^{(t)} || \alpha_c)\pi_{i,t}(\alpha_c) \tag{12}$$

Possibilities in practice are then very attractive, as one could possibly use previous scoring of a student as prior information on his class membership for the next progress test a few months later.

Based on the experience of the analyses in this thesis, some recommendations for future research can also be given. As has been stated in the method section 2.4, the decision on which restrictions to use for the latent class models could benefit from using a pseudo likelihood algorithm for choosing optimal class size in LCA. Another possibility to extent this research is to develop scoring methods for the three class solution, based on the three different conditional probabilities to belong to one of the classes given an answer pattern. Finally, a good practical test to find out whether theory(Baldiga, 2012) on penalty options also applies in the medical process test could be performed. It is stated that for tests with a penalty option could bias the test score for some sub-populations. Therefore, it would be good to perform an experiment in which students answer the medical test without a penalty option and compare the results to the medical test with the penalty option.

Bibliography

[1] AdaPT. (2008). *Projectsvoorstel AdaPT: Adaptieve voortgangstoetsing in Toetsing en Toetsgestuurd Leren.* Maastricht: Surf Foundation.

[2] Baldiga, K. (2012). *Gender differences in willingness to guess.* (Doctoral dissertation), Harvard University, 1-54.

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

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

[5] Dayton, C. M. (1998). *Latent class scaling analysis.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 107-126. Thousand Oaks, CA: Sage.

[6] Eggen, T. J. (2008). *Adaptive testing and item banking.* In J. Hartig, E. Klieme, & D. Leutner (Eds.), Assesment of Competencies in Educational Contexts (199-217). Cambridge, USA: Hogrefe & Huber Publishers.

[7] Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum Associates.

[8] Hagenaars, J., & McCutcheon, A. (Eds.). (2002). *Applied latent class analysis models.* New York: Cambridge University Press.

[9] 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.

[10] Jiao, H., Macready, G., Liu, J., & Cho, Y. (2012). A mixture Rasch model based computerized adaptive test for latent class identification. *Applied Psychological Measurement, 36,* 469-493.

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

[12] Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

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

[14] McCutcheon, A. L. (1987). *Latent Class Analysis.* Beverly Hills: Sage Publications.

[15] 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.

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

[17] Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

[18] Van der Linden, W. J., & Glas, C. A. (2000). *Computerized Adaptive Testing: Theory and practice.* Dordrecht: Kluwer Academic Publishers.

[19] 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,* 273-291.

[20] Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD User's Manual.* Boston: Statistical Innovations

[21] Vermunt, J. K. (in press). *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.

[22] Wainer, H. (2000). *Computerized Adaptive Testing: A primer.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

[24] Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361-375.

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

Appendix

Table 1: 2-Class Latent Class model probabilities to answer correct in 2005

item nr	**Class 1**	**Class 2**	**item nr**	**Class 1**	**Class 2**
1	0.6044	0.1939	17	0.6307	0.1356
2	0.6091	0.0515	18	0.4744	0.1342
3	0.8516	0.3766	19	0.5365	0.1694
4	0.4994	0.0466	20	0.3392	0.0246
5	0.5558	0.1845	21	0.5615	0.1194
6	0.1958	0.0698	22	0.6939	0.0994
7	0.6437	0.3004	23	0.4281	0.0212
8	0.4062	0.1810	24	0.4526	0.2083
9	0.3122	0.0054	25	0.6913	0.2815
10	0.7577	0.1271	26	0.8255	0.2137
11	0.9885	0.5291	27	0.4201	0.1298
12	0.3087	0.0316	28	0.6982	0.1671
13	0.8967	0.4591	29	0.5346	0.1333
14	0.6128	0.1087	30	0.7798	0.6093
15	0.4898	0.0521	31	0.4729	0.2209
16	0.551	0.1596	32	0.7896	0.2298

Table 2: 3-Class Latent Class model probabilities to answer correct in 2005

item nr	**Class 1**	**Class 2**	**Class 3**	**item nr**	**Class 1**	**Class 2**	**Class 3**
1	0.6459	0.4357	0.1012	17	0.6894	0.4126	0.0313
2	0.7309	0.2370	0.0261	18	0.5410	0.2981	0.0612
3	0.8706	0.7045	0.2455	19	0.6697	0.2419	0.1444
4	0.6157	0.1762	0.0296	20	0.4395	0.0996	0.0056
5	0.6289	0.3533	0.1176	21	0.6489	0.3036	0.0620
6	0.2270	0.1393	0.0213	22	0.8554	0.2661	0.0694
7	0.7020	0.4851	0.2131	23	0.5444	0.1280	0.0025
8	0.4328	0.3429	0.0850	24	0.5090	0.2993	0.1798
9	0.4054	0.0735	0.0000	25	0.7058	0.6028	0.1192
10	0.8616	0.4030	0.0571	26	0.8837	0.5685	0.0884
11	0.9960	0.9179	0.3208	27	0.4945	0.2467	0.0734
12	0.3630	0.1353	0.0123	28	0.8092	0.3682	0.1170
13	0.9182	0.8174	0.2565	29	0.6074	0.3191	0.0654
14	0.7032	0.3211	0.0533	30	0.7816	0.7670	0.5152
15	0.5939	0.1911	0.0291	31	0.5048	0.3617	0.1651
16	0.5869	0.4105	0.0488	32	0.8822	0.4731	0.1695