Research Master's programme: Methodology and Statistics in the Behavioural and Social Sciences
Utrecht University, the Netherlands

MSc Thesis *Hilde Matthea van Dijk*
TITLE: Measuring and Predicting Pupils' Progress in Special Education
May 2010

Supervisors:
Frans H. Kamphuis
Dr. Dave J. Hessen

Preferred journal of publication: Journal of Educational Measurement
Word count: 5111

Measuring and predicting pupils' progress in special education

Mathilde Huisman - van Dijk,[*] Frans H. Kamphuis[†] & Dr. Dave J. Hessen[‡]

May 19, 2010

Abstract

In regular primary education, LVS tests are used to measure and predict pupils' progress in math ability (θ). In special education, due to measurement error and a distribution of scores that is different from regular education, both measuring and predicting progress is difficult. A solution to the problem in measurement is investigated in the field of computerized adaptive testing (CAT). Using a simulation study, two CAT item selection mechanisms have been tested, for three different ability levels. Results suggest that using CAT mechanisms, accurately estimating a pupil's ability is possible, as long as the items are suitable for the ability levels. Overall, allthough the selection mechanisms are different, both mechanisms show equal performance in accuracy of measurment. For the problem of predicting progress, a distribution for special education is calculated and tested. The results suggest that, due to regression to the mean, the special education distribution does not predict accurately for pupils with θ values far from the population mean. Using Growth Mixture Modeling (GMM) multiple distributions have been defined to solve this problem. Using multiple distributions for pupils with different ability levels, the problem of regression to the mean is less severe.

Keywords: CAT, GMM, LVS, Primary Education, IRT, OPLM

[*] Department of Methods and Statistics, Utrecht University

[†] Department of Psychometrics, Cito institute of educational measurement, Arnhem

[‡] Department of Methods and Statistics, Utrecht University

1 Introduction

The Dutch National Institute for Educational Measurement (in Dutch called Cito) has developed a system that makes it possible to monitor pupils' progress in several skills throughout their entire primary school career. This system is called the Student Monitor System (in Dutch called LVS). For each skill the system consists of a set of tests that are administered twice a year, for each year of primary education. The student's scores on those tests are converted into one scale, so the results can be measured in a continuous and longitudinal way.

The scale that is studied in this paper is the math scale. Each test in the math scale is used to measure math ability, beginning in grade three, continuing into grade eight. However not all tests are used to measure the same part of the math scale. The tests in grade three are easier than the tests in grade five, so the tests in grade three cover a certain part of the total math scale, and the tests in grade five cover another part. Each test is used to measure the same latent construct, and each test covers a piece of the total scale, reaching from grade three on to grade eight. This way a student's progress can be measured in a longitudinal way, and can be monitored and compared throughout his school career. Besides the monitoring, by using growth models the system can be used to predict a pupils future results (Kamphuis & Engelen, 1993).

The LVS is not only administered in regular primary education but also in special primary education. The way progress is measured and predicted in regular education will be explained in section 2 of this paper. Using the LVS measuring progress and predicting future outcomes is in special education not as straightforward as in regular primary education. Problems occur in both measuring and predicting ability in special education.

Measuring a special education pupil's ability is complex. The LVS tests are administered such that a pupil in regular education should be able to make the LVS test corresponding to the grade that pupil is in (Kamphuis & Moelands, 2000). In special education this is a problem. For example, a special education pupil in grade six does not necessarily have the ability corresponding to the level that can be expected from a pupil in grade six in regular education. On average pupils in special education are behind on the LVS schedule and there is a great amount of variability in the ability levels of pupils within one grade in special education. This way it is uncertain what test should be administered in order to make an accurate measure of a pupils ability at that time. When the administered test is too difficult or too easy for a pupil the test becomes uninformative, and no precise estimate can be made about its ability level (Eggen, 2004). Therefore a solution has to be found to make measuring ability in special education more accurate.

Also predicting a pupil's progress in special education is problematic. The program used to predict progress in the LVS system uses a regression model based upon the means and variances of the population of children in regular education. However, the distribution of scores in special education is different from regular education. When using the regular education distribution for special education, two problems occur. The first problem is that the means of regular education are too high for special education. This results in unrealistic predictions due to regression to the mean. Because the means in regular education are higher than the means in special education, the ability levels of pupils in special education are being overestimated in the predictions. The second problem is that the variance of scores in special education is much larger than in regular education. This results in inaccurate predictions and wide confidence intervals around the predictions.

In 2006 a pilot study has been performed under ten special education schools in the Netherlands in order to create a model, that can be used to predict pupils' progress in special education. This model was based upon the means and variances of a special education population. The results of the study show that using that model, it is possible to make more realistic predictions. However, the accuracy of this model has not been tested yet. Also because of the large variance in special education, one single model may not be enough to cover the whole population of pupils in special education. Therefore the accuracy of the special education model should be tested, and when necessary multiple subpopulations have to be defined and for each subpopulation a prediction model has to be estimated, increasing the accuracy of prediction in special education (CITO, 2007).

The aim of this study is to improve the accuracy of both measurement and predictions in special education. Therefore first a possible solution to the problem of inaccurate and uninformative measurement is presented. A solution to this problem may lie in using the technique of Computerized Adaptive Testing (CAT), which is a technique in which tests are individualized such that each pupil makes a test that is appropriate for its own ability level. In order to investigate whether using CAT can solve the problem of inaccurate and uninformative testing, simulation studies have been done. In these simulations several CAT applications are investigated in order to find a CAT method that is appropriate for the pupils in special education and can solve the problem. Then a solution to the problem of inaccurate predictions is presented. A population model is estimated for special education and using several simulations the accuracy of this model in several applications will be tested. Also Growth Mixture Modeling (GMM) will be used to search for underlying latent classes within the population of special education pupils, in order to define multiple subpopulation models that can make predictions more accurate.

This paper is structured as follows. In section two, the measurement model used for the construction of the math scale is explained. Also the growth model used for prediction is explained in this section. In section three, the framework of CAT is explained and the simulations investigating accuracy of measurement in special education are done and discussed. In section four, the simulations investigating the accuracy of predictions in special education are done and discussed. Also a way to improve the accuracy of predictions is presented. In section five, a discussion and conclusion is presented.

2 Measurement and prediction

The scale used in this thesis is the Cito math scale, which is measured by a series of tests administered throughout a pupils primary school career. Each year a great amount of schools use the Cito math tests and send the results of their pupils back to Cito. The first test is usually administered halfway grade three (denoted by M3), then every half a year a test is administered, continuing to the end of grade eight (denoted by E8). Tests that are administered halfway a grade are specified by the character M, tests that are administered at the end of a grade are specified with the character E. The total scale ranges from 0 to 153. These numbers correspond to θ values, representing the level on the latent trait of math ability.

In this section the measurement model used for the construction of the math scale is described. Also the growth model, used to predict a person's skill on a future time point is explained.

2.1 Measurement model

The items in the math tests are calibrated using Item Response Theory (IRT). In IRT the item response function is crucial, this is the function that gives the probability to a correct answer to the item, given a persons ability. In IRT several models to calculate this probability are possible. The model that is used to estimate the item parameters in the math scale, is called the One Parameter Logistic Model (OPLM) (Verhelst, Glas & Verstralen, 1995). In the OPLM model, the item response function is given by

$$P(X = 1 \mid \theta) = f_i(\theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}$$

where θ represents a person's math ability, β_i is the difficulty of item i and a_i is the discrimination parameter of item i (Verhelst, 1993). X is ar random variable with possible values 1 and 0. This function gives the probability of a correct answer ($X = 1$) to item i with discrimination parameter a_i and difficulty parameter β_i as a function of θ. All items in a test have their own β value, estimated by the model described above. All items together form a scale ranked from the easiest item to the most difficult item. By the probabilities of making an item correctly by a person with abillity θ, this scale can be used to estimate a pupil's θ. For a complete scale the likelihood is then the probability of answer pattern $\boldsymbol{x}$ given θ assuming local independence, and can be denoted by

$$P(\boldsymbol{x} \mid \theta) = \prod_{i=1}^{n} P_i(\theta)^{x_i}[1 - P_i(\theta)]^{1-x_i}.$$

In the OPLM model a_i is a chosen constant so it is not a parameter that has to be estimated. Fixing the discrimination parameter a_i makes it possible to estimate the difficulty parameters in the OPLM model with Conditional Maximum Likelihood (CML). In CML for every item the total number of responses in a particular response category will be used for the estimation of the parameters of the item. This number is a sufficient statistic for the parameter. CML estimates are obtained by conditioning on sufficient statistics for the person parameters, which are the persons' sum scores, weighted by the discrimination parameters. Item parameters are estimated by equating the sufficient statistics to their expected values, conditional on the frequency distribution of the persons' scores (Verhelst, Glas & Verstralen, 1995).

2.2 Growth model

The longitudinal design of the math scale

The LVS scales are designed such that the same latent construct is measured through time. As stated in the introduction, each test covers a part of the total math scale, so the data can be analyzed and monitored in a longitudinal design. To be able to analyze the test results in a longitudinal design, either a static or a dynamic approach can be used.

In the static approach the test results are analyzed in cross sections. This means that for each individual the point estimates of θ for each measurement occasion are compared with each other, and the development through time can be monitored. In this case a pupil's point estimate of θ at a certain time point can be estimated using the Weighted Maximum Likelihood estimater (WML), introduced by Warm (1989). This estimator is also called the Warm estimator and it is the value of θ that maximizes the the likelihood function, weighted by the square root of the testinformationfunction. The Warm estimator is given by

$$\hat{\theta}_{wml,t} = \text{Max}P(s_{w,t} \mid \theta_t)\sqrt{I(\theta)}$$

where $\hat{\theta}_t$ is the estimated ability for time point t, $s_{w,t} = \Sigma a_i x_{it}$, which is the sum of a pupil's correct items x_{it} for $i = 1 \ldots n$, weighted by the discrimination parameter a_i at time point t, and $\sqrt{I(\theta)}$ is the square root of the test information function.

However, when measuring progress this way a problem occurs. When a test is administered a certain amount of measurement error is involved, resulting in an irregular and unpredictable progress profile. For example, when the point estimate of θ on a certain time point is higher than the estimate on the foregoing timepoint it can be due to progress, but it can also be due to measurement error.

In order to make better estimations of θ, it can be assumed that each pupil is part of a certain reference population. This population has a certain distribution of scores through time with a mean $\boldsymbol{\mu}$ and a variance-covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is a matrix describing the relations between the scores on each timepoint. In this case the estimator of θ is called the 'Expected A Posteriori' estimator or EAP. The estimator is given by

$$\tilde{\theta}_t = E(\theta_t \mid s_{w,t}, \mu_t, \sigma_t^2) = \frac{\int \theta_t P(s_{w,t} \mid \theta_t) P(\theta_t) d\theta_t}{\int P(s_{w,t} \mid \theta_t) P(\theta_t) d\theta_t}$$

and its variance by

$$Var_t(\theta_t \mid s_{w,t}, \mu_t, \sigma_t^2) = \frac{\int (\theta_t - \hat{\theta}_t)^2 P(s_{w,t} \mid \theta_t) P(\theta_t) d\theta_t}{\int P(s_{w,t} \mid \theta_t) P(\theta_t) d\theta_t}$$

where $P(s_{w,t} \mid \theta_t)$ is again a pupil's probability of weighted score $s_{w,t}$, given its θ, $P(\theta_t)$ is the distribution of θ in the population with the mean μ and variance σ^2, and θ is the ability of the population at time point t (Oud & van Blokland-Vogelenzang, 1993).

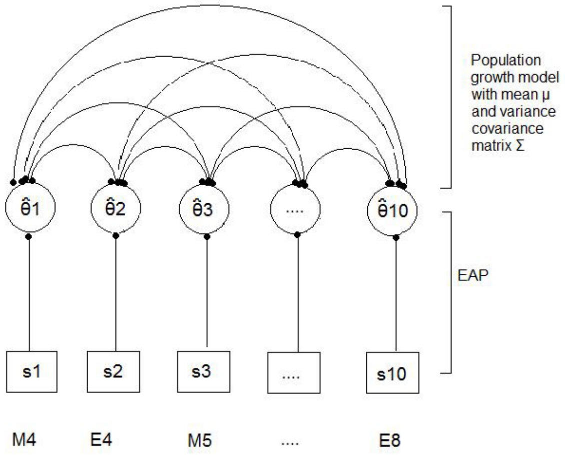

Figure 1: Population growth model

Estimation using the population growth model

In order to monitor a pupil's progress in a longitudinal way, also a dynamic aproach can be used. In this approach, using the relations between the measurement points in the population a growth model is estimated. The relations in this model can also be used to predict a pupils progress on future time points (Kamphuis & Engelen, 1993).

The growth model is a regression model of wich the structural part can be described with

$$\boldsymbol{\theta} = \boldsymbol{\alpha} + \boldsymbol{B}\hat{\boldsymbol{\theta}} + \boldsymbol{\zeta}$$

for $t = 1 \ldots n$, where $\boldsymbol{\theta}$ is a $T \times 1$ random vector representing the prediction of θ for time points t=1 $\ldots T$, $\boldsymbol{\alpha}$ is a constant $T \times 1$ vector representing te intercepts, $\boldsymbol{B}$ is a $T \times T$ lower triangular matrix representing the covariance between the measurements, and $\boldsymbol{\zeta}$ is a $T \times 1$ zero mean disturbance vector uncorrelated with θ and with diagonal covariance matrix (Joreskog & sorbom, 1996). T represents the total number of timepoints, so in this case T= 1 $\ldots 10$. In reduced form, the model is

$$\boldsymbol{\theta} = (\boldsymbol{I} - \boldsymbol{B})^{-1}(\boldsymbol{\alpha} + \boldsymbol{\zeta}).$$

Using this equation $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be calculated. Figure 1 gives a graphical representation of the model.

Using this model the progress profile can be smoothened and the estimations of a pupil's true ability at a certain time point can be made more reliable. θ_t is estimated conditional on all other θ_t values. This is done as follows. First the population distribution is defined. Using the computer program MULTI the scores of the reference population are used to estimate a multivariate population distribution with a mean $\boldsymbol{\mu}$ and a variance covariance matrix $\boldsymbol{\Sigma}$ (Dempster, Laird & Rubin, 1977).

Then the following steps are taken:

Step 1: Estimate θ_1 given s_1, μ_1, σ_1^2 (This is the EAP estimation for θ_1).
Step 2: Insert $\tilde{\theta}_1$ in the growth model to estimate θ_2 given $\theta_1, s_2, \mu_2, \sigma_2^2$.
Step 3: Insert $\tilde{\theta}_1$ and $\tilde{\theta}_2$ in the growth model to estimate θ_3 given $\theta_1, \theta_2, s_3, \mu_3, \sigma_3^2$.
Step 4: Continue untill all θ_t values are estimated conditional on all foregoing $\tilde{\theta}_t$ values.

Now using these steps, the first measurement occasions are used to estimate the last measurement occasions, however to smoothen the progress profile the same steps are taken backwards. So $\tilde{\theta}_9$ will be smoothened conditional on $\tilde{\theta}_{10}$, $\tilde{\theta}_8$ will be smoothed conditional on $\tilde{\theta}_{10}$ and $\tilde{\theta}_9$, and so forth, untill all θ_t values are estimated conditional on all $\tilde{\theta}_t$ values. With each step also the estimation error covariance, belonging to the estimates of θ is calculated. In the end a new distribution is created with the expecatation $E(\theta \mid s_w, \mu, \Sigma)$ and covariance $\text{cov}(\theta \mid s_w, \mu, \Sigma)$.

Predictions using the population growth model

The population growth model can also be used to predict a pupil's progress on future time points (Kamphuis & Engelen, 1993). By using the model in combination with the first few scores of a pupil, future scores can be estimated forming a pupil's distribution of ability θ through time. For example when only two test are administered and a prediction of the fourth time point must be made, θ_1 is estimated by the EAP estimator, θ_2 is estimated using the EAP estimator and $\tilde{\theta}_1$. But for θ_3 the score is not known, so it is estimated conditionally on $\tilde{\theta}_1$ and $\tilde{\theta}_2$, and θ_4 is estimated by $\tilde{\theta}_3$,$\tilde{\theta}_2$ and $\tilde{\theta}_1$.

3 Measurement in special education

As stated in the introduction assigning the appropriate tests to pupils in special education is problematic. Figure 2 shows measurement error functions of four of the math tests as well as the probability density of θ in both regular and special education, for θ values between 0 and 100. The M3 test is calibrated on pupils of mid grade three, the E3 test is calibrated on pupils of end grade three, M4 corresponds to mid grade four and E4 corresponds to the end of grade four. All tests are calibrated on pupils in regular primary education. As can be seen in the figure, the measurement error increases instantly when administering it to a population with an ability level that is too high or too low for the test. For example, the bold solid line represents the measurement error of the test administered at the end of grade four (test E4). The measurement error is smallest when made by pupils with a θ value of 60 on the scale. However, when looking at the probability density of pupils in special education at the end of grade four (represented by the bold dotted line), it can be seen that on average the pupils have a θ value of 45. The line also shows that the variance around that mean is relatively large compared to regular education, and that the tails of the special education density are thicker. So for many special education pupils in E4, the test administered in E4 does not measure well because of measurement error.

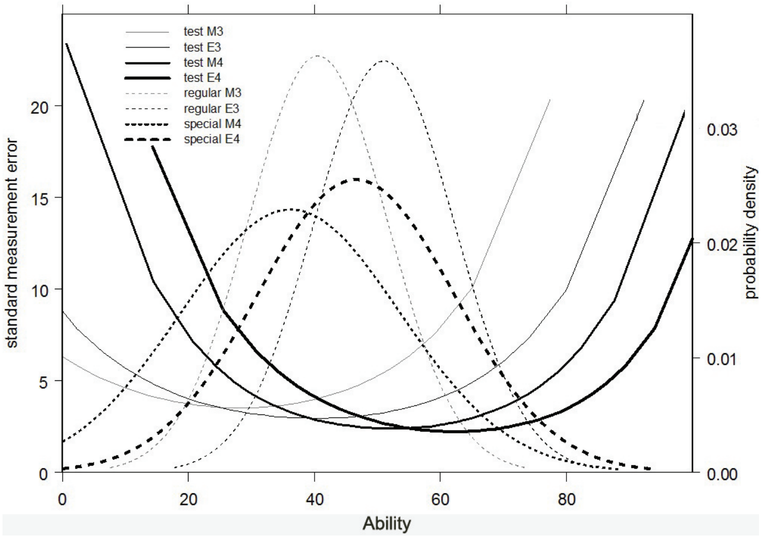

Figure 2: Measurement error for M3 to E4

Simply administering an easier test for pupils in special education is not a solution. Because of the large variance in the distribution of θ in special education it is uncertain where a pupil's true θ lies on the scale. The bold dotted line in figure 1 shows that a large range of θ values are clustered within one grade. Assigning the test that fits a pupil's latent ability level is thus a challenge which makes measuring progress in special education pupils difficult. For example, when a special education pupil scores low at an E4 test, this could mean that its ability level is low, but it could also mean that the test does not fit its ability level properly.

In this section a solution for this problem is investigated within the framework of Computerized Adaptive Testing (CAT). In the first subsection the framework of CAT is explained and in the second subsection it is investigated whether using CAT as a solution to the problem of measurement in special education is reasonable.

3.1 Computerized Adaptive Testing

When the measurement error of a test is too large, the number of items should be increased in order to reduce the error. However, the LVS Math tests already have a large amount of items, a test with 120 items is not exceptional. Many special education pupils cope with problems in their concentration and get distracted when the test is too long. Increasing the number of items would in this case likely increase the measurement error instead of decreasing it. In order to decrease the measurement error in special education a test is needed that is as short as possible, using items with difficulty parameters close to a pupil's true ability level.

Computerized Adaptive testing can be a solution. In CAT the construction and administration of the test is computerized and individualized. Each pupil gets a test that is appropriate for its own ability level. Using an item selection algorithm the items that are to be administered to a pupil are selected, conditional on whether the forgoing item was made correctly or not. The items are contained in an item bank, and from each item in the item bank the psychometric properties are considered to be known. The items are ranked on the same ability scale as the pupil is measured on. When a pupil answers an item correctly, there is a large probability that the item lies lower on the θ scale than the θ value of the pupil, and a more difficult item is selected. When that item is made incorrectly an easier item is selected. This proces continues until convergence is reached (Eggen, 2004). Making use of items this way has advantages. The selection of items can be made broader without increasing the actual length of the test. This means that the test can measure beyond the range of ability levels of the tests that are used in regular education. For example, the items of test E3, M4 and E4 can be put in one CAT itembank so the CAT test covers a broad part of the θ scale.

There are many possible item selection mechanisms and each mechanism uses its own item selection criteria. For special education a mechanism is needed that is able to make fast and accurate measurement possible. In the next subsection two selection mechanism are compared in order to investigate how accurate CAT measures, and which approach suits special education best.

3.2 Accuracy of CAT measurement

In order to investigate the accuracy of the measurements using CAT, a simulation study has been performed, comparing the accuracy of CAT at three different θ values, and two different item selection mechanisms. Because of the large variance in the θ distribution in special eduction, the method is tested on a pupil with a low θ value (θ=30), a pupil with a mean θ value (θ=50) and a pupil with a high θ value (θ=80). For all tests the same itembank is used, with items that are usually in the LVS tests M3 and E3. In the simulation each pupil makes two tests. The two tests use different item selection meachnisms. In order to test the accuracy of the mechanisms each test is made 1000 times.

The math scale as it is used in practice ranges from θ values of 0 to 153. However, the items are calibrated on another scale, and are then transformed into the θ scale. Because the difficulty parameters in the itembank are on the untransformed scale, the untransformed values of the θ scale are used in this simulation. The untransformed values of θ=30, θ=50 and θ=80 are respectively -.349, .246 and 1.138.

Selection mechanism

The selection mechanisms used in this simulation are the Fisher Information Selection Criteria (FISC) and the Expected Shannon Information Selection Criteria (ESISC). FISC uses the most common selection mechanism: it maximizes the Fisher information given the estimated θ distribution conditional on the foregoing answers. For example if the first item is made, θ is calculated, and the item will be selected for which the pupil has 50% chance to answer it correctly. If the next answer is correct, a new θ value is calculated, and again the item for which the pupil has 50% chance to make it correctly is selected. This continues until a certain predestined stopping point, in this case after 25 items. The probability of making the item correctly does not necessarily have to be 50%.The amount of information also depends on the item discrimination parameter.

ESISC uses another selection mechanism, in which the tails of the total distribution of θ are cut off with each item. In order to select an item the predictive distribution is used, based on the Expected Shannon Information (Bernardo & Adrian, 1994). The Shannon information is the average amount of information the observer has gained after receiving a correct answer to an item (Grunwald & Vitanyi, 2004). In ESISC the expected distribution of the shannon information is maximized, and used for item selection. In practice relatively easy items are alternated with relatively difficult items. With each answer the θ distribution for the pupil becomes more narrow, this continues untill the distribution of θ becomes stable.

Figure 3 shows which items are selected in both mechanisms for a pupil with θ=30. The bold line represents the difficulty of the selected items. The horizontal line in both graphs represents the θ value of the pupil, and the dotted lines represent the estimated θ value and its 80% confidence interval. As can be seen in both cases the items are selected such that around item 15 the pupil has to make items that are close to its true θ value, and thus give much information. In both cases θ is slightly overestimated, but the true value lies within the confidence interval.

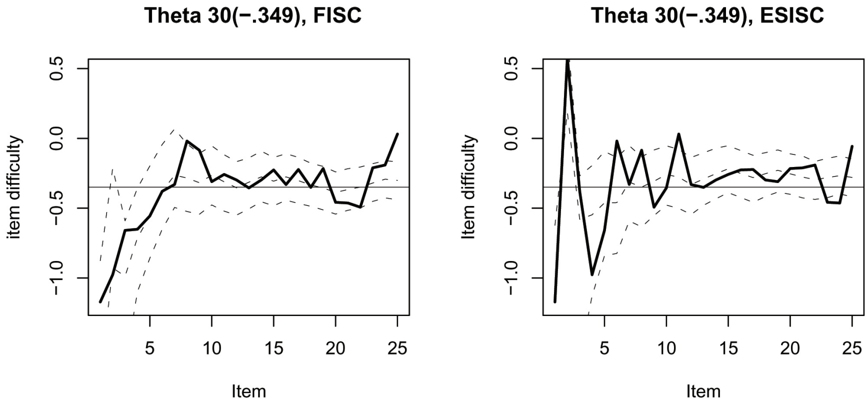

Figure 3: Selection mechanisms FISC and ESISC for θ = 30

Accuracy of measurement

To be able to select the CAT item selection mechanism that suits special education best, the mean accuracy of the measurements have to be investigated. Table 1 shows the mean estimated θ values for each item and the mean range of the 80% confidence interval, for the θ values 30, 50 and 80, and for both selection mechanisms. The results show that both mechanisms do about equally well with low θ values. With $\theta = 30$ both mechanisms on average slightly overestimate θ with estimates of around -0.355, whereas the untransformed value of θ is -0.349. Also the mean 80% confidence interval shows that both mechanisms are equally accurate, both intervals have about the same range, with values of respectively .267 and .275 for FISC and ESISC. With average θ values the same conclusion can be drawn. Both mechanisms show with $\theta = 50$ that they both slighty underestimate θ, with .239 for FISC and .241 with ESISC, and both reach equal accuracy values, with .280 for FISC and .281 for ESISC. With high θ values also both mechanisms do equally well. Both mechanisms underestimate θ. With mean estimates of around 1.107 and 1.115, ESISC is on average closer to the true θ value than FISC, but shows a higher 80% confidence interval, with values of respectively 1.029 for FISC and 1.125 for ESISC.

It can be concluded that both mechanisms are able to accurately estimate θ when the true θ value is around 30 or 50. On average the results are less accurate with θ values 80, and are less certain. A closer look at the data shows that on average pupils with a θ of 80 make 23.67 of the 25 items correctly. This makes measuring accurately very difficult because of a ceiling effect. This means that when almost all items are made correctly, not much information is given except for the fact that the pupil probably has a high θ value. The exact level however cannot be estimated accurately. Figure 4 shows the items that are selected for a pupil with $\theta = 80$. The bold black line represents the item difficulties, and the horizontal line represents the true θ value. The figure shows that the items in the itembank are not difficult enough to give information of the correct θ value of the pupil. For all items the probability of making the item correctly is likely to be large.

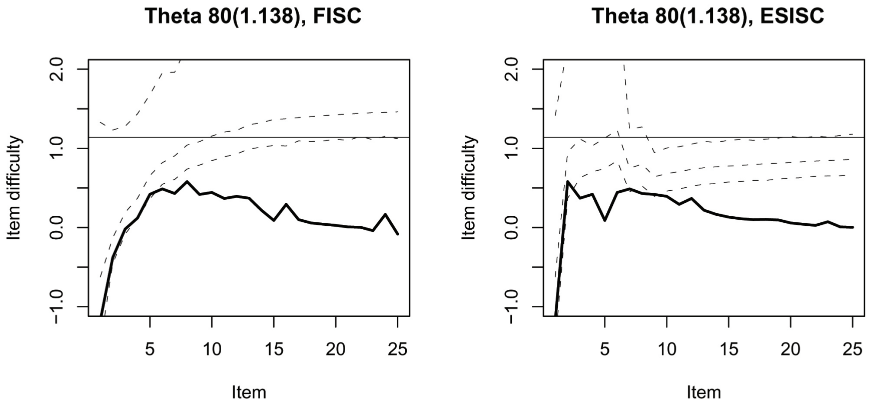

Figure 4: Selection mechanisms FISC and ESISC for $\theta = 80$

Table 1: Results of the simulation with θ values of 30(-0.349), 50(.246) and 80(1.138)

	θ=30, FISC		θ=30, ESISC		θ=50, FISC		θ=50, ESISC		θ=80, FISC		θ=80, ESISC	
	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
	est	80%	est	80%	est	80%	est	80 %	est	80%	est	80%
Item	θ	Range	θ	range	θ	Range	θ	Range	θ	Range	θ	Range
1	-0.809	2.852	-0.809	2.854	-0.683	2.858	0.683	2.853	-0.627	2.854	-0.627	2.850
2	-0.456	1.535	-0.084	2.222	-0.192	1.711	0.309	2.084	-0.133	1.751	0.824	1.984
3	-0.370	0.884	-0.382	1.315	0.067	1.225	0.227	1.277	-0.185	1.389	0.762	1.800
4	-0.361	0.687	-0.326	0.953	0.160	0.953	0.209	0.976	0.365	1.313	0.890	1.857
5	-0.362	0.585	-0.335	0.742	0.209	0.690	0.247	0.743	0.640	1.243	0.942	1.787
6	-0.362	0.529	-0.354	0.632	0.223	0.590	0.249	0.630	0.932	1.261	0.984	1.679
7	-0.362	0.488	-0.353	0.559	0.229	0.528	0.245	0.562	0.847	1.242	1.017	1.673
8	-0.362	0.459	-0.351	0.511	0.230	0.483	0.243	0.506	0.932	1.322	1.040	1.674
9	-0.361	0.431	-0.353	0.476	0.235	0.452	0.243	0.471	0.964	1.232	1.056	1.582
10	-0.359	0.409	-0.355	0.448	0.235	0.428	0.242	0.441	1.006	1.244	1.069	1.528
11	-0.361	0.391	-0.356	0.425	0.236	0.406	0.241	0.417	1.030	1.208	1.077	1.473
12	-0.359	0.375	-0.355	0.404	0.238	0.388	0.242	0.396	1.040	1.166	1.086	1.427
13	-0.360	0.361	-0.356	0.386	0.240	0.373	0.243	0.380	1.066	1.270	1.090	1.388
14	-0.359	0.348	-0.357	0.372	0.238	0.360	0.243	0.366	1.072	1.223	1.099	1.361
15	-0.358	0.337	-0.357	0.358	0.240	0.349	0.243	0.354	1.080	1.220	1.099	1.330
16	-0.358	0.327	-0.358	0.346	0.237	0.337	0.243	0.344	1.093	1.219	1.103	1.299
17	-0.357	0.318	-0.359	0.335	0.237	0.328	0.241	0.333	1.097	1.192	1.103	1.265
18	-0.358	0.309	-0.358	0.324	0.238	0.319	0.240	0.324	1.098	1.163	1.106	1.245
19	-0.357	0.301	-0.358	0.316	0.238	0.311	0.239	0.315	1.100	1.137	1.107	1.226
20	-0.359	0.294	-0.357	0.307	0.239	0.305	0.241	0.308	1.103	1.114	1.109	1.202
21	-0.356	0.288	-0.357	0.299	0.239	0.299	0.241	0.301	1.106	1.099	1.111	1.189
22	-0.356	0.282	-0.355	0.292	0.238	0.293	0.240	0.295	1.107	1.082	1.113	1.171
23	-0.355	0.276	-0.356	0.286	0.239	0.289	0.239	0.295	1.107	1.059	1.115	1.155
24	-0.355	0.271	-0.356	0.280	0.239	0.284	0.240	0.285	1.108	1.046	1.116	1.142
25	-0.355	0.267	-0.355	0.275	0.239	0.280	0.241	0.281	1.107	1.029	1.115	1.125

In order to investigate the reliability of the mechanisms the 80% coverage rate for each test is calculated. For FISC with a θ of 30, in 20.5% of the cases the true θ value falls outside the 80% confidence interval. That is slightly more than the 80% confidence interval would allow. For ESISC with a θ of 30, in 19,3% of the cases θ falls outside the interval, which is reasonable. The coverage rate for the $\theta = 50$ simulations are with both mechanisms .20, which is reasonable. For FISC with a θ of 80, in only 9.7% of the cases θ falls outside the interval and for ESISC with θ 80, in 20.2% of the cases θ falls outside the confidence interval.

In general these results suggest that using CAT for measurement in special education is a reasonable option. Using short tests a good indication of a pupils ability can be estimated. However, as in the normal LVS tests, CAT has its limitations. The selection of items can be made more broad, but then still it has to fit ones true θ value. If the true θ does not fit the item bank, accurate measurement is still difficult.

4 Predictions in special education

When a child enters special education, usually not much is known about the ability of this pupil and about what is to be expected from it. A prediction of his progress is however wanted at a early stage of its school career, so teachers know what to aim for (Clijssen et al., 2009). As stated in the introduction, predicting progress in special education is complicated. A large amount of background variables influence the distribution of the data, resulting in a large variability of scores, and a large variability in possible growth compared to regular education. When the regular education population distribution is used in the growth model described in section 2, due to regression to the mean the predictions for special eduction will be unrealistic and inaccurate.

In this section, using the program MULTI a special education population distribution is estimated. Then, using several simulations the accuracy of predictions made using this new distribution is investigated. A dataset is created and using the information from different numbers of time points, predictions are made of the ability level θ at the tenth time point using the population growth model described in section 2.

The accuracy is investigated for several different situations. Schools cannot always provide all information about the tests that are administered and the items that are made correctly. In the ideal situation, all information of the administered test is available, and the weighted score (denoted by s_w) can be calculated. This means that it is known what test is made, what items de pupil answered correctly and which are answered incorrectly. Using the item discrimination parameter a the scores x on the items are weighted, so $s_w = \Sigma ax$. The s_w is a sufficient statistic for θ, this means that s_w can be used to estimate θ, by inserting it in the EAP estimator described in section 2.

However in practice sometimes only Σx (the number of correct answers on the test) is provided, without the information of which items are answered correctly, and which are answered incorrectly. Then only unweighted scores (denoted by s_{uw}) are available for the prediction. In this case for each item the chance of scoring the item correctly are included in the process of calculating θ.

In some cases there is no information available about the test that was made at all. Then the item parameters are not known and cannot be included in the growthmodel, and cannot be included in the estimates and predictions. In that case only 'Warm estimations' of θ, (denoted by $\hat{\theta}_{wml}$) are available. When item parameters are not known, the scores that are used to make the predictions are slightly biased. There is an amount of measurement error involved, which adds up to the error in the predictions.

4.1 Accuracy of predictions with s_w, s_{uw} and $\hat{\theta}_{wml}$

Creating data

To be able to test the accuracy of the predictions simulation studies have been done. A complete dataset is created with measurements available on all time points, of which $\hat{\theta}$ can be calculated. Then measurement points are deleted from the data so that predictions can be made and tested against the known values from the complete dataset. For the simulation of the parameters, realistic values have been used, seen in a study after growth modeling in special education.

The dataset is created as follows. First the parameters θ for each time point have been simulated for each person in the sample. Then using these values response patterns on a series of tests are created measuring math ability through time. This way a sample of test scores are created for each person in the sample, for each time point, resulting in a complete dataset, with which $\tilde{\theta}$ can be calculated.

To make sure the measurement error is as small as possible, the test that fits the pupil's ability level is assigned, such that each pupil has a 60% chance to pass the test.

The dataset contains 800 pupils with measurements on 10 time points, beginning from the middle of grade 4 (M4), untill the end of grade 8 (E8). Table 2 shows the descriptives of the complete dataset. As can be seen in the table, the means of the scores range from 36.19 on the first measurement to 86.45 on the tenth measurement. These are realistic values in special education whereas in regular education the scores are on average about twenty points higher on the scale and the standard deviation is smaller in regular education.

Table 2: Descriptives.

Timepoint	Grade	Mean $\hat{\theta}$	SD
1	M4	36.19	17.41
2	E4	46.59	15.64
3	M5	52.09	16.00
4	E5	57.94	15.48
5	M6	64.93	14.05
6	E6	69.70	14.33
7	M7	74.07	13.94
8	E7	78.56	14.34
9	M8	82.73	15.03
10	E8	86.45	15.36

Special education population distribution

As described in section 2, first the population distribution is estimated. Table 3 shows the correlations between the θ values, on the different time points in the population distribution as it is estimated by MULTI. As can be seen the correlations between the values are high. This shows that the scores on the different time points can be used well for the prediction of future scores. The correlations are highest between scores on adjacent time points. This is to be expected because the scores on the different time points are assumed to measure the same construct on a continuous scale. Scores on adjacent timepoints will then be more alike than scores further apart on the scale. Using only these first measures for the prediction of scores up to the tenth measurement will probably produce uncertain predictions.

Table 3: Correlation matrix of the outcome distribution.

Timepoint	1	2	3	4	5	6	7	8	9	10
1	1.00	-	-	-	-	-	-	-	-	-
2	.939	1.00	-	-	-	-	-	-	-	-
3	.884	.856	1.00	-	-	-	-	-	-	-
4	.833	.775	.938	1.00	-	-	-	-	-	-
5	.820	.826	.849	.861	1.00	-	-	-	-	-
6	.772	.731	.880	.865	.893	1.00	-	-	-	-
7	.764	.742	.864	.852	.865	.935	1.00	-	-	-
8	.769	.770	.876	.850	.859	.917	.944	1.00	-	-
9	.682	.736	.861	.868	.860	.916	.927	.932	1.00	-
10	.668	.718	.863	.818	.865	.924	.909	.923	.968	1.00

Accuracy of predictions using s_w

To be able to test the accuracy of the predictions, measurements on several time points are deleted, and then predicted using the population growth model described in section 2. First only the tenth measurement was deleted, and a prediction was made using the information of the other available time points. Then the tenth and the ninth time points were deleted, then the eighth to the tenth, and so forth. Each time a prediction of the tenth measurement was made using the information of the remaining time points. The predicted tenth measurement can then be compared with the true tenth measurement point.

Table 4 shows the means, variances and standard deviations from the MULTI predictions of the tenth measurement point. As can be seen in the table, the earlier the predictions are started, the less information is used and the less accurate the prediction becomes. What is striking is that the less information is used, the more the predictions tend to underestimate $\hat{\theta}$ on average, where on the most precise predictions $\hat{\theta}$ is slightly overestimated. The turning point is between the predictions using 1 to 5 timepoints and 1 to 6 timepoints.

Table 4: Means and SD's for predictions of θ using s_w.

Measurements used	Mean of predictions	Difference with true mean	Mean SD
Timepoint 1 to 9	86.690	0.240	3.221
Timepoint 1 to 8	86.728	0.278	3.988
Timepoint 1 to 7	86.717	0.267	4.302
Timepoint 1 to 6	86.759	0.309	4.634
Timepoint 1 to 5	86.475	0.025	5.935
Timepoint 1 to 4	86.443	-0.007	7.304
Timepoint 1 to 3	86.350	-0.100	7.527
Timepoint 1 to 2	85.295	-1.155	10.875
Only Timepoint 1	85.951	-0.499	11.613

Accuracy of predictions using s_{uw}

All given results so far, concern s_w values, which is the ideal situation. In this subsection it is investigated whether using s_{uw} for prediction results in a difference in the prediction or not, and what the effect is on the accuracy of the prediction.

Table 5 shows the summarized results of the predictions using weighted and unweighted scores and the differences between them. As can be seen the results using unweighted scores are very similar to the predictions using weighted scores. The differences in the predicted means are over all slightly higher with unweighted scores than in the weighted scores, except for the predictions using 1 to 8 time points and from 1 to 5 time points. However in all cases the difference is very small, with .013 points of difference at the most in the prediction using 1 to 2 time points. Also in the standard deviations there is just a small difference with at the most .038 points. What is striking here is that standard deviations are more similar in the last two predictions, being the one with 1 to 2 time points and the one with only 1 time point. Here the standard deviation differs only .008 and .007 points whereas in the other predictions the difference is around .045.

Table 5: Means and SD's for predictions of θ using s_w and suw.

Timepoints	Mean			mean SD		
used	(s_w)	(s_{uw})	(diff)	(s_w)	(s_{uw})	(diff)
1 to 9	86.690	86.691	.001	3.221	3.262	.041
1 to 8	86.728	86.727	-.001	3.988	4.029	.041
1 to 7	86.717	86.719	.002	4.302	4.347	.045
1 to 6	86.759	86.763	.004	4.634	4.684	.050
1 to 5	86.475	86.469	-.006	5.935	5.979	.044
1 to 4	86.443	86.447	.004	7.304	7.336	.032
1 to 3	86.350	86.352	.002	7.527	7.565	.038
1 to 2	85.295	85.308	.013	10.875	10.883	.008
Only 1	85.951	85.954	.003	11.613	11.620	.007

Accuracy of predictions using $\hat{\theta}_{wml}$

In stead of creating test scores on each time point for each student, $\hat{\theta}_{wml}$ has been calculated from the weighted scores of each time point. Using the same growth model as in the previous simulations predictions for each tenth time point have been made.

Table 6 shows the results of the analysis using $\hat{\theta}_{wml}$ as input. The results show that using $\hat{\theta}_{wml}$ for predictions does not produce any flaws compared to using the weighted and unweighted test scores. The mean standard deviations are only slightly higher and the predictions are very similar to the results shown in table 5.

Table 6: Means and SD's for predictions of θ using $\hat{\theta}_{wml}$.

Measurements used	Mean of predictions	Difference with true mean	Mean SD
Timepoint 1 to 9	86.596	.146	3.217
Timepoint 1 to 8	86.569	.119	4.004
Timepoint 1 to 7	86.556	.106	4.318
Timepoint 1 to 6	86.600	.150	4.649
Timepoint 1 to 5	86.322	-.217	5.943
Timepoint 1 to 4	86.318	-.132	7.314
Timepoint 1 to 3	86.241	-.209	7.532
Timepoint 1 to 2	85.342	-1.108	10.876
Only timepoint1	86.156	-.294	11.627

4.2 Reducing error using multiple population distributions

Regression to the mean

As seen in the analysis in the previous subsection, all given results give about the same predictions when time points are taken away from the data. However, the results also show that the accuracy in prediction decreases fast when using less information. In all results it can be seen that the interval around the mean prediction becomes very large and a lot of uncertainty has to be taken for granted.

In the used prediction method the mean of the prior population distribution on each time point is very important. A consequence of this is that when predictions are made of measurements of more than one time point further in time than the current measurement, the prediction will be drawn towards the mean of the population. This means for pupils that score low relatively to the mean, their predicted score will be overestimated, and for pupils that score high relatively to the mean, their predictions will be underestimated. Because of the large variance in the scores in special education this problem becomes severe, especially for long term predictions.

A solution to the problem of regression to the mean may be in defining multiple population distributions. So far, it is assumed that all pupils are part of the same distribution. That means that the same model is used for the prediction of scores of all pupils, irrespectively of their level of ability. The predictions may be made more accurate when different distributions are assumed for pupils with different levels of ability, or with different growth speed. When the distribution of scores in special education is split up into several smaller distributions, the variance around each mean will be smaller, and the problem of regression to the mean will be less severe.

Growth Mixture Modeling

In order to define multiple distributions, Growth Mixture Modeling (GMM) is used. GMM assumes that the data consist of a mixture of multiple distributions in stead of one multivariate normal distribution. In GMM a growth model is fitted to the data, estimating an intercept and a slope for a defined number of latent classes (Muthen, 2004).

The model used in this study can be written as $Y_t = i + bt + bt^2 + e$, where Y_t represents the outcome for the measurement at timepoint t, i represents the intercept, bt represents the slope of time and e represents the error. Because growth in math ability is likely to be slightly curved a quadratic term is added to the model, represented by bt^2. This model is estimated for each latent class separately.

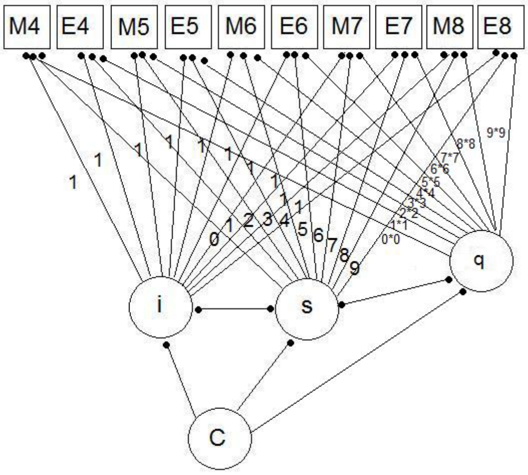

Figure 5: Growth Mixture Model

Figure 2 shows a graphical representation of the model. The loadings of the measurements on the intercept are fixed at 1 in order to be able to calculate the intercept, and the loadings of the measurements on the slope increasing from 0 to 9, representing a linear growth curve. The quadratic term is represented by the 'q', of which the loadings are fixed to the squares of 0 to 9 (Kline, 2005; Muthen & Muthen, 1998-2007).

Estimating the number of latent classes

In order to decide the number of latent classes that should be defined, several analyses have been performed. 4000 θ values have been simulated, having the same properties as the data used in the analysis in the foregoing subsection. The growth model shown in Figure 5 has been fitted on the data, first on four classes, then on three, and so forth. With each analysis a Vuong-Lo-Mendell-Rubin Likelihood Ratio Test has been performed in order to decide whether defining a model with that particular number of classes is a significant improvement on the model with one class less (Muthen & Muthen, 1998-2007). Table 7 shows the results of these tests. As can be seen, deciding on 2 latent classes is the best solution on these data.

Table 7: Vuong-Lo-Mendell-Rubin Likelihood Ratio Test results

Number of latent classes	H0 loglikelihood value	2 times log likelihood difference	Difference in number of parameters	Mean	SD	p
2 vs 1	-139902.669	101.149	4	5.350	4.810	.000
3 vs 2	-139852.094	6.570	4	4.781	6.326	.268
4 vs 3	-139848.809	18.555	4	6.831	10.469	.106

Results per latent class

Table 8 shows the results of the GMM analysis with two latent classes. For each class the estimated intercept, slope and quadratic term are shown as well as the percentage of pupils that are classified to be part of that latent class. The results show that 88,4% of the pupils are part of the second latent class. On average these pupils start with a θ value of 39.13, and grow on average 8.11 points between measurements 1 and 2. However, the quadratic term shows that the growth decreases each measurement with a quadratic term of .29. Most pupils in special education will likely belong to this class. 11.6% of the pupils belong to the first latent class, starting at a lower θ value, with an intercept of 11.62. On average they grow slightly faster than the largest group with a slope of 10.07 between measurements 1 and 2, however this slope is also decreasing faster with a quadratic term of -.44.

Table 8: Latent classes

Class	Intercept	Slope	Quadratic term	% of pupils
1	11.619	10.065	-0.441	11.6
2	39.125	8.110	-0.291	88.4

Figure 6 shows a graphical representation of the results shown in Table 8. Both the estimated values and the observed values are represented in the graph. As can be seen the estimated and the observed scores lie close to each other. The quadratic growth curve is thus a good representation of the growth in the data. Also the figure shows that despite the differences in slope and quadratic term, the lines seem to be parallel. The line of the first latent class is only slightly more curved.

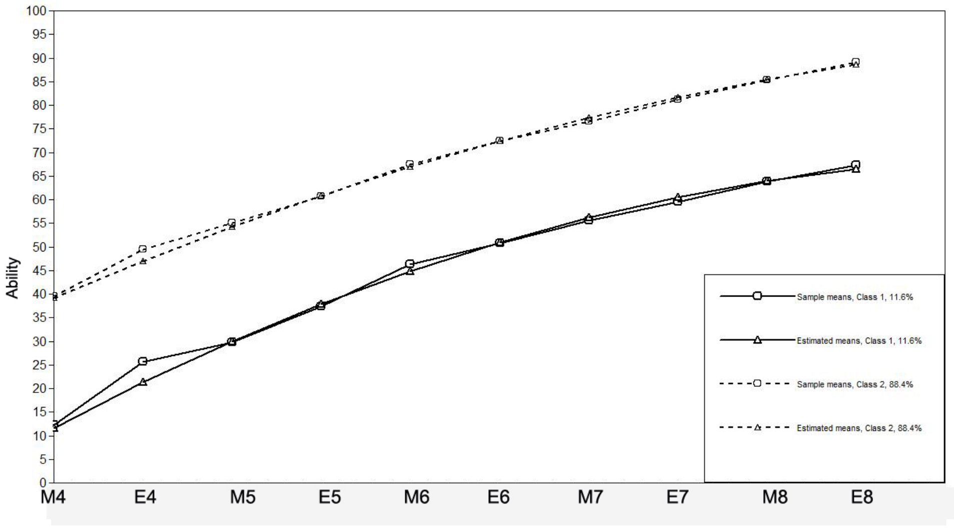

Figure 6: Estimated and observed means of both latent classes

Assigning pupils to distributions

Defining multiple distributions creates a new problem. In order to make an individual prediction, a decision should be made for each pupil to which latent class it belongs. The majority of the pupils will be classified to be part of the second latent class and for pupils with scores around the means of both distributions the decision will not be difficult. However it is more challenging when dealing with pupils that score in between both distributions.

Table 9 shows the probabilities for each class by membership of each class. The columns represent the actual latent classes, and rows represent the most likely latent class. As can be seen in the table, a pupil that is classified to be in class 1 has a probability of .26 that it should belong to class 2. This 26% is likely the part of the distribution of whitch the results lie between both classes. Because of the great amount of pupils in class 2, the probability of actually belonging to class 1 is only 7 percent. For these pupils it is useful to take the chances to belong to the other class into account.

Table 9: Probabilities for most likely class membership

Class	1	2
1	0.738	0.262
2	0.066	0.934

Accuracy of predictions using multiple population distributions

Table 10 shows the means and standard deviations on all time points for both latent classes. The results in the table show that the standard deviations remain relatively large. They are only slightly smaller than the standard deviations of the total distribution as it is seen in Table 2. This is not a surprise since the total distribution is split up into only two parts, with one part representing 88.4% of the pupils. However, the problem of regression to the mean was the largest with very low θ values, deviating strongly from the mean of the population, because the number of pupils with very low scores was relatively large. Because the most extreme deviating pupils will be part of a class with a lower mean, the problem of regression to the mean will be smaller, and a more accurate prediction can be made for each pupil.

Table 10: Mean scores and standard deviations on all time points for both latent classes

Time point	Class 1 Mean	SD	Class 2 Mean	SD
M4	12.188	15.23	39.531	15.09
E4	25.561	13.83	49.417	13.64
M5	29.829	14.08	54.988	13.61
E5	37.354	14.02	60.636	13.28
M6	46.208	12.64	67.493	12.38
E6	50.759	12.76	72.341	12.54
M7	55.759	12.41	76.574	12.24
E7	59.517	12.84	81.195	12.43
M8	63.735	13.51	85.267	13.17
E8	67.195	13.90	89.031	13.55

5 Discussion and conclusion

The aim of this study was to improve the accuracy of measurement and predictions in special education. First a solution to the problem of inaccurate and uninformative measurement was presented within the framework of Computerized Adaptive Testing (CAT). Several applications were tested in order to find a CAT method that fits the pupils in special education. The results show that using CAT in special education would be resonable. With low θ values both tested methods behave properly, giving accurate θ values and small confidence intervals. Also a good estimation can be made in a short test, which has an advantage for children in special education. However, as in the normal LVS tests, CAT has its limitations. The selection of items can be made more broad, but then still it has to fit ones true θ value. If the true θ does not fit the item bank, accurate measurement is still difficult.

With respect to the accuracy in predictions in special education, a dataset is created from which measurement points were deleted. A distribution was estimated and used for the prediction of the deleted tenth time point. In general the results have shown that the accuracy of the prediction decreases fast when more time points are deleted from the data. The estimations of θ on the tenth measurement occasions are on average close to the real mean of the data, but the standard deviation is on average very large. The simulation was done for several situations, including weighted scores s_w, unweighted scores s_{uw} and Warm estimations $\hat{\theta}_{wml}$. The results have shown that using s_{uw} and $\hat{\theta}_{wml}$ as input does not result in any flaws in the predictions compared to input using s_w. The difference with the predictions using s_w are only small and also the standard deviation increases only little. This is a useful result in the practice of eduction, because it means that relatively easy programs can be used as productively as sophisticated and complex programs including the measurement model and all item parameters. However these results have to be interpreted with care. The data used for these simulations are created under the OPLM model. The results of these simulations can not necessarily be generalized to tests that are calibrated under another model. Also in the the simulations, the test results were simulated such that the test fitted the pupil's ability level well. In practice this is not alway the case. These simulations can be repeated on real data, in order to investigate the accuracy of predictions in less optimal circumstances.

In long term predictions the problem of regression to the mean occurred, making the predictions inaccurate. To make the predictions more accurate a Growth Mixture Model is fitted to the data to search for underlying latent classes. The results show that the total distribution can be split up into two groups with different intercepts and growth curves. The largest part of the population of special eduction pupils are part of the same group, but a small group with scores far beneath the mean of the population form a different group. When these groups are taken into account when predicting pupils' progress the accuracy in prediction can be improved, especially for the pupils with low scores. However the given results are still exploratory. The model has to be tested on real data in order to check whether this structrure truly exists in the general population of special education pupils in the Netherlands. To investigate this, further study on this particular topic is needed.

References

Bernardo, J. M., & Adrian, F. M. (1994). *Bayesian theory.* Smith John Wiley & Sons, Chichester.

CITO. (2007). *Stand van zaken onderzoek leerrendementsverwachting per september 2007.* (Unpublished manuscript)

Clijsen, A., Pieterse, E., Spaans, G., & Visser, J. (2009). *Werken vanuit een ontwikkelingsperspectief in het speciaal basisonderwijs, naar een gezamelijk kader.* www.sbowerkverband.nl.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1-38.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing.* Cito, Arnhem.

Grunwald, P., & Vitanyi, P. (s. d.). *Shannon information and kolmogorov complexity.* (Manuscript)

Joreskog, K. G., & Sorbom, D. (1996). *Lisrel 8: User's reference guide.* Chicago: Scientific Software International.

Kamphuis, F. H., & Engelen, R. J. H. (1993). Het meten van veranderingen. In *Psychometrie in de praktijk.* Cito, Arnhem.

Kamphuis, F. H., & Moelands, F. (2000). A student monitoring system. *Educational Measurement: Issues and Practice*, *19*, 28-30.

Kline, R. B. (2005). *Principles and practice of structural equation modeling, 2nd edition.* Guilford, New York.

Muthen, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In *Handbook of quantitative methodology for the social sciences.* Newbury Park, CA: Sage Publications.

Muthen, L. K., & Muthen, B. O. (1998-2007). *Mplus users guide. fifth edition.* Los Angeles, CA: Muthen & Muthen.

Oud, J. H. L., & Blokland-Vogelenzang, R. A. W. van. (1993). *Advances in longitudinal and multivariate analysis in the behavioral sciences.* ITS, Nijmegen.

Verhelst, N. D. (1993). Itemresponstheorie. In *Psychometrie in de praktijk.* Cito, Arnhem.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *One-parameter logistic model oplm.* Cito, Arnhem.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.