Computers & Education 57 (2011) 1813–1824

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation

S. Klinkenberg*, M. Straatemeier, H.L.J. van der Maas

Department of Psychology, University of Amsterdam, Roeterstraat 15, Room A522, 1018 WB Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 26 November 2010
Received in revised form
8 February 2011
Accepted 8 February 2011

Keywords:
IRT
CAT
CAP
Computer adaptive practice
Serious gaming
Progress monitoring
Item calibration

ABSTRACT

In this paper we present a model for computerized adaptive practice and monitoring. This model is used in the Maths Garden, a web-based monitoring system, which includes a challenging web environment for children to practice arithmetic. Using a new item response model based on the Elo (1978) rating system and an explicit scoring rule, estimates of the ability of persons and the difficulty of items are updated with every answered item, allowing for on the fly item calibration. In the scoring rule both accuracy and response time are accounted for. Items are sampled with a mean success probability of .75, making the tasks challenging yet not too difficult. In a period of ten months our sample of 3648 children completed over 3.5 million arithmetic problems. The children completed about 33% of these problems outside school hours. Results show better measurement precision, high validity and reliability, high pupil satisfaction, and many interesting options for monitoring progress, diagnosing errors and analyzing development.

1. Introduction

In this paper we present a computerized adaptive practice (CAP) system for monitoring arithmetic in primary education: the Maths Garden. The Maths Garden[1] is a web-based computer adaptive practice and monitoring system based on weekly measurements. In recent years the Maths ability of Dutch students has been widely debated. This is mainly due to the results of the National Periodical Education Polls (PPON). These results show that few children reach the required Maths level at the end of their primary education (Kraemer, Janssen, Van der Schoot, & Hemker, 2005). Based on these findings a parliamentary inquiry into Dutch education was initiated. Both the committee "Dijsselbloem" (2008) and the expert group "Doorlopende Leerlijnen" (2008) recommended several improvements to the Dutch education system in general and Maths education in particular. Recommendations included the provision of more time to practice and maintain basic Maths skills, more efficient and effective measurement in education and the use of these measurements results to improve the ability of individual students, the classroom and education in general. These recommendations are also supported by Fullan (2006) who claimed that acting on data is critical for learning from experience.

1.1. Combining practice and measurement

In the light of these recommendations we propose to combine practice and measurement in a playful manner using computerized educational games. We expect that in the near future children will increasingly use mini computers and handheld devices to do their daily exercises in arithmetic, spelling, and other subjects. The use of computers have two main advantages. First, the input can be analyzed automatically and feedback can be given immediately which will free teachers from checking and correcting the children's exercise books. The recorded and automatically analyzed data can provide teachers with detailed information on children's progress and the errors they make. Teachers can use this information to optimize individual instruction. The information concerning the child's progress and abilities, which is accumulated over time, may ultimately obviate the need to conduct tests and examinations. Second, by using computers it is

* Corresponding author. Tel.: +31 20 525 6721.
E-mail address: s.klinkenberg@uva.nl (S. Klinkenberg).

[1] The Dutch version is called: Rekentuin.nl and is used by more than 150 schools. The English version MathsGarden.com started in the summer of 2010.

0360-1315/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2011.02.003

possible to let children practice at their individual ability level. Research on the development of expertise performance has shown that people do improve their performance considerably if they regularly do specific exercises that are adjusted to their ability level and include immediate feedback. In the development of the Maths Garden we follow these ideas developed in sports and expertise training, especially the idea of deliberate practice (Ericsson, 2006, pp. 683–703).

1.2. Three problems of CAT

To implement individualized practice, we apply the technique of computer adaptive testing (Van der Linden & Glas, 2000; Wainer et al., 2000). Computer adaptive testing (CAT) is based on item response theory (IRT). This theory consists of statistical models that relate item responses to the (latent) abilities that the items measure (Lord & Novick, 1968). A large collection of item response models is available, but these are all basically variations on the simplest model, i.e., the one-parameter logistic (1PL) model or Rasch model (Rasch, 1960). In the Rasch model the probability of a correct or affirmative answer is a logistic function of the difference between the ability of the subject and the difficulty of the item. In the two-parameter logistic model, the difference is weighted by an item discrimination parameter, which has a high value when an item discriminates well between low and high ability subjects. Item response models can be used for equating tests, to detect and study differential item functioning (bias) and to develop computer adaptive tests (Van der Linden and Hambelton, 1997). The idea of CAT is to determine the ability level of a person dynamically. In CAT, item administration depends on the subject's previous responses. If the preceding item is answered correctly (incorrectly), a more (less) difficult item is presented. Hence, each person is presented a test tailored to his or her ability. Using CAT, test length can be shortened up to 50% (Eggen & Verschoor, 2006). Originally, CAT was developed for measurement only. Our aim to combine practice and measurement raises several novel issues. We distinguish the following three issues.

First, in standard CAT the parameters of the items, especially the difficulty, have to be known in advance to test adaptively. Items therefore have to be "pre-calibrated" before they can be used in real test situations. This means that a large representative sample of the population has to have answered the items in the item bank to provide the information for item calibration. The difficulty of the items is determined using the data of this sample. This method is obviously time consuming and costly, especially as the calibration has to be carried out repeatedly (e.g. every few years) to acquire accurate norm referenced item parameters.

Second, CAT operates most effectively if the difficulty level of administered items equals the ability estimate of the person. The probability of answering such items correctly is .5. However, for most children and many adults the success rate associated with a .5 probability is experienced as discouraging. Research by Eggen & Verschoor (2006) showed that increasing this probability to above .7 greatly reduces measurement precision. Given a .7 probability, more items need to be administered to obtain an accurate estimate of person ability. This requirement reduces the efficiency of computer adaptive testing.

The third problem concerns a testing problem that applies to psychological and educational measurement in general, namely, the trade-off between speed and accuracy. Without explicit instructions, participants in tests and experiments are free to balance speed and accuracy as they wish. Consequently the trade-off between speed and accuracy can be a source of large individual differences. The current solution in psychometrics (Van der Linden, 2007) and experimental psychology (Ratcliff & Rouder, 1998; Vandekerckhove & Tuerlinckx, 2008) is to estimate person parameters involved in this trade-off on the basis of the data. However, this procedure requires large amounts of high quality data.

1.3. New CAT

We developed an extended CAT approach to solve these problems. This Computer Adaptive Practice (CAP) system provides the basis of the Maths Garden. The CAP system includes the following two innovations. First, we have applied a new estimation method based on the Elo (1978) rating system (ERS) developed for chess competitions. The ERS allows for on the fly estimation of item difficulty and person ability parameters. With this method, pre-testing is no longer required. Second, we have used an explicit scoring rule for speed and accuracy, which is known to the subject during the test. Inclusion of speed in the scoring has the advantage that we acquire more information about ability. Research by Van der Maas & Wagenmakers (2005) showed that in the response to easy chess items there is a strong negative relation between response time and ability. Subjects tend to answer easy items correctly, but more advanced subjects answer them more quickly. Third, by integrating response time in the estimation of ability, we can decrease the difficulty of administered items with less loss of measurement precision than noted by Eggen & Verschoor (2006). In addition we expect the higher success rate to increase the motivation of children during the test. In the Method section we describe the Maths Garden, the Elo algorithm and the new scoring rule in more detail. In the results section of this paper we test the working of Maths Garden. We present evidence for high validity and reliability of ability and difficulty estimation, the motivational value of the Maths Garden and its usefulness as a diagnostic and monitoring instrument.

2. Methods

2.1. Participants

A total of 35 primary schools, eight remedial teachers and 32 families participated in this study, comprising of $N = 3648$ active participants. Also 334 aspiring kindergarten pupils joined the Maths Garden. In the time period from August 2008 to early June 2009 more than 3.5 million arithmetic problems were answered in our sample. In addition to the responses we registered the gender, age and grade of the participants. Table 1 shows the mean age with standard deviation and the amount of children for each grade.

2.2. Materials

The main measurement tool used in this study is the web-based practice and monitoring system we developed: Maths Garden. The student interface consists of a garden containing distinct flowerbeds, representing, among others, the four domains: addition, subtraction, multiplication and division (Fig. 1a) on which we focus in this paper. The size of the flowers represents the Maths ability of the student. By clicking on a flowerbed the Maths game is started for a specific domain and the student can start playing.

Table 1
Age, gender and *N* per grade.

Grade	Age category	N	Age $\bar{x}(\sigma^2)$	%M(F)
kindergarten	4-5	103	4.32(0.51)	50.49(49.51)
kindergarten	5-6	231	5.45(0.51)	47.19(52.81)
1	6-7	529	6.61(0.51)	53.50(46.50)
2	7-8	681	7.69(0.54)	55.21(44.79)
3	8-9	526	8.68(0.79)	47.91(52.09)
4	9-10	513	9.70(0.61)	47.24(52.76)
5	10-11	574	10.79(0.60)	49.48(50.52)
6	11-12	416	11.80(0.57)	50(50)
Secondary Education	12 <	75	13.33(3.94)	64(36)

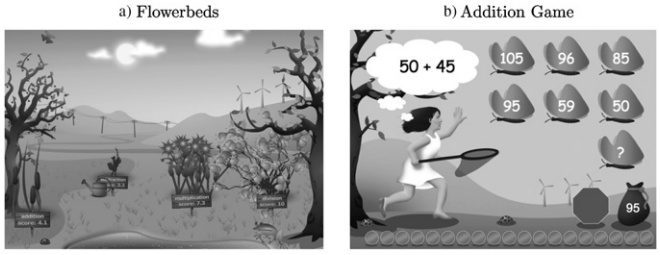

Fig. 1. The main Maths Garden interface and an addition item.

The visual interface of the Maths task consists of a Maths question, six answer options, a coin bag, a question mark, a stop sign, and an elapsing coin bar which indicates the time left on the item (Fig. 1b). The game rules are intuitive and therefore only require minimal description on the website. Students gain points (coins), displayed at the bottom of the game interface, by answering items correctly and lose coins when answering incorrectly. With every item a total amount of twenty coins, corresponding to the maximum time in seconds, can be won or lost. Every second one coin disappears. The remaining coins are added to the coin bag if the item has been solved correctly and are subtracted if solved incorrectly. If the time limit has expired or the question mark has been clicked, no coins are lost or won. The rationale of this scoring rule is explained in the psychometrics section. A session consists of fifteen items after which the Maths game terminates and the student is returned to his or her garden. The flowers will start growing according to the progression that has been made. Students are motivated by two reward systems. Good performance is rewarded by growing flowers and virtual coins[2]. The Maths Garden website contains a dedicated area, the prize cabinet, where virtual prizes can be bought with the earned coins. Another way students are motivated to continue playing in the Maths Garden is to have the flowerbeds wither if the student does not play. Withering worsens over time and can only be undone by completing a new session of 15 items.

The four domains, addition, subtraction, multiplication and division, contain 738, 723, 659 and 664 items, respectively. The items in the four domains cover the curriculum in primary education. They vary from easy (e.g., 3 + 4 with response options: 7, 8, 6, 1, 9 and 12) to difficult (e.g., 7,34 + 311,4 with response options: 318,74; 318,38; 318,47; 317,74; 319,74 and 318,34). The response options are selected to be informative distracters. Variables measured by the task are response time, the given answer, the correctness (0, 1) and a timestamp at administration.

We studied the validity of the data by comparing the ability estimate, measured with the Maths Garden, with students' scores on the Maths tests from the pupil monitoring system (Janssen & Engelen, 2002) of the National Institute for Educational Measurements (CITO). In the CITO monitoring system Maths tests are administered twice a year from mid grade 1 until mid grade 6. These tests assess the knowledge and skills that are being taught in these grades. The tests contain both open-ended and forced-choice items. Students' scores on the Maths test (the total of correct answers) are transformed to a score on a norm-referenced general Maths ability scale. This allows one to compare students' scores from different grades using one scale.

2.3. Psychometrics

2.3.1. Elo rating system

In chess the Elo (1978) rating system (ERS) is used to estimate the relative ability of a player. The ERS is a dynamic paired comparison model which is mathematically closely related to the Rasch IRT model (Batchelder & Bershad, 1979). Initially chess players are given a provisional ability rating θ which is incrementally updated (see equation (1)) based on match results (in chess 0, 0.5 and 1, for loss, draw and win outcomes). The updated ability estimate $\widehat{\theta}$ (sgnified by the hat) depends on the weighted difference in match result S and expected match result $E(S)$. The expected match result is a function of the difference between the ability estimates of both player j and k preceding the match and expresses the probability of winning (see equation (2)):

$$\begin{aligned}\widehat{\theta}_j &= \theta_j + K(S_j - E(S_j)),\\ \widehat{\theta}_k &= \theta_k + K(S_k - E(S_k)),\end{aligned} \tag{1}$$

[2] Because of the adaptive nature of the test, every student has roughly the same percentage correct. Hence the number of coins won reflects only how often a student plays and not his arithmetic level.

$$E(S_j) = \frac{1}{1 + 10^{(\theta_j - \theta_k)/400}}. \quad (2)$$

The K factor in equation (1) weights the impact of the deviation from expectation on the new ability estimate. This value essentially determines the rate at which θ can change over matches. In the standard ERS the K factor is constant. Glickman (1995) argued that not all ability ratings are estimated accurately by the ERS update function (eq. (1)). Inaccuracies mostly occur when players are new or have not played for an extended period of time, resulting in much uncertainty in their ability rating θ. Glickman proposed to let the K factor reflect the uncertainty in ability estimates by making it a function of time and playing frequency. If there is little uncertainty, the K factor for recent and frequent players will be low. If there is much uncertainty the K factor will be high.

2.3.2. Computer adaptive practice

Our suggestion for creating an on the fly item calibrating and computer adaptive practice (CAP) system is to replace one player in the Elo system by an item.[3] Solving an item correctly is interpreted as winning the match against the item. The updating function in equation (1) can be rewritten to equation (3) for updating player and item ratings:

$$\begin{aligned} \widehat{\theta}_j &= \theta_j + K_j(S_{ij} - E(S_{ij})), \\ \widehat{\beta}_i &= \beta_i + K_i(E(S_{ij}) - S_{ij}), \end{aligned} \quad (3)$$

where β_i is the difficulty estimate of the item and S_{ij} and $E(S_{ij})$ are the score and expected probability of winning for person j on item i. Following Glickman, the K factor in our CAP system is a function of the rating uncertainty U of the player and the item (eq. (4)):

$$\begin{aligned} K_j &= K(1 + K_+U_j - K_-U_i), \\ K_i &= K(1 + K_+U_i - K_-U_j), \end{aligned} \quad (4)$$

where $K = 0.0075$ is the default value when there is no uncertainty and $K_+ = 4$ and $K_- = 0.5$ are the weights for the rating uncertainty for person j and item i. These values determine the rate at which θ and β can change following each item response. These values have been determined through extensive simulations.

The uncertainty U depends on both recency and frequency. Equation (5) combines these opposite effects on uncertainty. We apply the same equation to items and players, with provisional uncertainty of $U = 1$ and $0 \leq U \leq 1$:

$$\widehat{U} = U - \frac{1}{40} + \frac{1}{30}D. \quad (5)$$

We assume that uncertainty for players and items decreases after every administration and increases with time. Therefore uncertainty reduces to zero after 40 administrations and conversely increases to the maximum of 1 after 30 days D of not playing.

2.3.3. High speed, high stakes

We incorporate speed by using the scoring rule (shown in eq. (6)) for speed and accuracy, which we call the high speed high stakes (HSHS) scoring rule (Maris & Van der Maas, Submitted for publication). This rule imposes a speed accuracy trade-off setting on the individual. Player j has to respond x in time t_{ij} before the time limit d_i for item i. The score S_{ij} is scaled by the discrimination parameter a_i:

$$S_{ij} = (2x_{ij} - 1)(a_i d_i - a_i t_{ij}). \quad (6)$$

In this scoring rule the stakes are high when the subject responds quickly. In case of a correct answer ($x_{ij} = 1$) the score equals the remaining time. In case of an incorrect answer ($x_{ij} = 0$) the remaining time is multiplied by -1. Thus a quick incorrect answer leads to a large negative score. This scoring rule is depicted in Fig. 2. The scoring rule is expected to minimize guessing by encouraging deliberate and thoughtful responses.

Maris & Van der Maas (Submitted for publication) derived an IRT model that conforms to the HSHS scoring rule. The expected score (eq. (7)) can be inferred from this model. $E(S_{ij})$ is based on the ability estimate of the person θ_j, the difficulty estimate of the item β_i, the time limit d_i and discrimination parameter a_i for that item. In the Maths Garden, we set $a_i = 1/d_i$, such that the effective discrimination equals that of the 1PL model:

$$E(S_{ij}) = a_i d_i \frac{e^{2a_i d_i(\theta_j - \beta_i)} + 1}{e^{2a_i d_i(\theta_j - \beta_i)} - 1} - \frac{1}{\theta_j - \beta_i}. \quad (7)$$

We use the HSHS score S_{ij} (eq. (6)) and the corresponding expected score $E(S_{ij})$ (eq. (7)) in our modified Elo update function (eq. (3)).

2.3.4. Item selection

Items are selected for which the mean probability of answering correctly is about .75. Repetition of the same items is restricted, by ensuring that items are reused only after 20 other items have been answered. A new target β_t is selected by using:

$$\beta_t = \widehat{\theta}_j + ln\frac{P}{1 - P}, \quad (8)$$

where probability P is randomly drawn from a normal distribution $\sim(0.75,0.1)$ and restricted such that $0.5 < P < 1$. For administration the nearest available item is selected by: $\min_i|\beta_i - \beta_t|$.

[3] This approach has, for many years, successfully been applied in an online chess testing system on the Chess Tactics Server (chess.emrald.net).

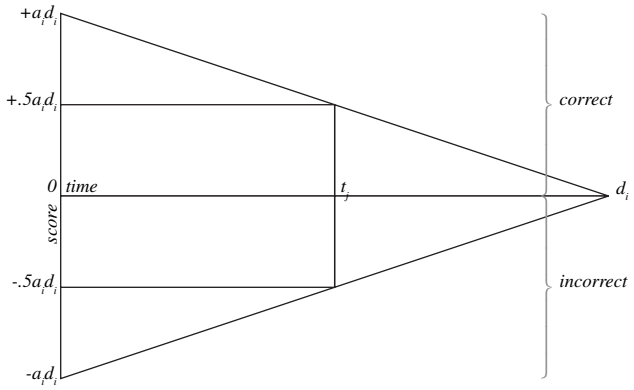

Fig. 2. High speed, high stakes scoring rule.

2.4. Procedure

Although the Maths Garden started out as a pilot project, only available to a limited amount of schools in the Netherlands, the website later on became available for a larger audience. In the pilot period the students received a login account and an instruction from their teacher. In this instruction, teachers explained the scoring rule of the games and students were told that they could click on the question mark if they did not know the answer. After this, students could start playing on their own. Teachers were told that the first two sessions should be played at school. After this, students were also allowed to play at home, they were also instructed to play by themselves. After the pilot period the Maths Garden also became available to remedial teachers and families. The remedial teachers and families were not instructed on the frequency of playing. The manuals on how to use the Maths Garden were all available on the website but the scoring rule of the games was not explicitly described to the children on the website.

3. Results

3.1. Measurement precision

To test whether the incorporation of response time in the estimation of ability allows us to lower the difficulty of administered items with less loss of measurement precision, we conducted a simulation study. We compared our results to those of Eggen and Verschoor (2006). In a simulation study, Eggen & Verschoor, showed[4] an increasing (negative) bias (Fig. 3: left) and a drop in measurement precision (Fig. 4: right) when selecting easy items in a standard CAT using the weighted maximum likelihood estimator (WML) and the one-parameter logistic (1PL) model. Average bias was computed by: $1/n\sum(\widehat{\theta}_i - \theta_i)$ and measurement precision was quantified by calculating the mean standard error of estimation $se(\widehat{\theta})$ using the information function for the 1PL model.

In our simulation we used the Elo update function to estimate ability and difficulty, utilizing: a) accuracy data with the 1PL model and b) accuracy and response time data using the HSHS model. As in the study by Eggen & Verschoor, our item bank consisted of 300 items with normally distributed $\beta \sim N(0,1)$ difficulties and we also sampled 4000 abilities from a normal distribution $\theta \sim N(0,1)$. The CAP algorithm starts with an item of intermediate difficulty $-0.5 < \beta < 0.5$ and terminates after 40 items. As a starting point for ability we selected a random ability from a normal distribution $\beta \sim N(0,1)$. We compared our Elo based HSHS model, at different desired success probabilities, to Eggen & Verschoor's 1PL model using standard CAT. Eggen & Verschoor investigated success probabilities up to .75.

With regard to bias it can be concluded that the Elo estimation method performs slightly worse with accuracy data only (Fig. 3: left: Elo+1PL), but outperforms Eggen & Verschoor 's standard CAT method, when RT's are included (Fig. 3: left: Elo + HSHS).

With regard to the standard error of estimation we also compared our two Elo methods to the theoretical maximum information for the 1PL model. We calculated the maximum information (Fig. 3: right: Max Info.) with equation (9):

$$se\left(\widehat{\theta}\right) = 1/\sqrt{Na_i^2 P_i(\theta)(1 - P_i(\theta))}, \tag{9}$$

where $\alpha = 1$ is the discrimination parameter, $N = 40$ is the number of items, and $P_i(\theta)$ is the desired probability correct. As can be seen in (Fig. 3: right: Max info.), when the probability of answering correctly assumes large values (x-axis), the theoretical minimum SE (eq. (9)) for the 1PL model increases exponentially (y-axis). For the standard error of estimation (Fig. 3: right) we calculated the standard deviation of the difference in simulated abilities θ and estimated abilities $\widehat{\theta}$. This method of calculation is simpler, yet comparable with the procedure used by Eggen & Verschoor for calculating the standard error of estimation.

The SE of the Elo estimation method using only accuracy data (Fig. 3: right: Elo+1PL) is largest for almost all probability levels. This is to be expected as this method is statistically inferior to the WML method used by Eggen & Verschoor. Up to the probability level of about .69 the SE using the HSHS Elo method (Fig. 3: right: Elo + HSHS) is larger than the SE found in the Eggen & Verschoor simulation. However, at higher probability levels, especially compared to our target of .75, the SE is considerably lower. At probability levels higher than about .78 the SE even drops below the theoretical maximum information (Fig. 3: left: Max info.) for the 1PL model. This demonstrates that incorporating response times results in much better measurement precision when using easy items.

[4] Table 1 in Eggen and Verschoor (2006).

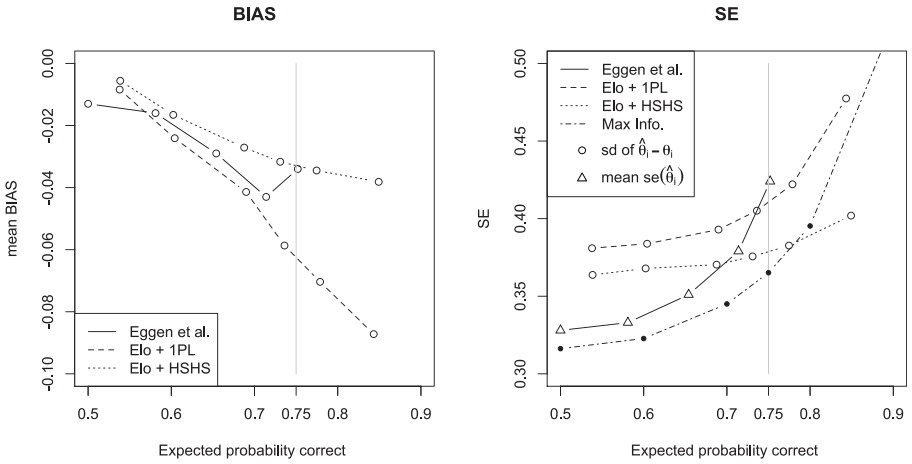

Fig 3. BIAS and SE for different computer adaptive methods at different values of the expected probability correct.

3.2. Validity

To assess the validity of the Maths Garden measurements, the ratings of the students were compared to their scores at the norm-referenced general Maths ability scale of the pupil monitoring systems of CITO (Janssen & Engelen, 2002). The correlations between these two measures, which serve as a measure of convergent validity, ranged from .78 to .84 for the four domains addition, subtraction, multiplication and division. These correlations were based on a subset of our sample. CITO scores where available for $N = 964$ participants. To put these correlations into perspective we looked at the correlation between two subsequent CITO scores. The correlation between CITO mid year and end of the year 2007–2008 was .95. This indicates that our correlations can be considered fairly high. Fig. 4 displays the relation between test scores. The numbers indicate the regression line for each grade.

We also studied the validity by comparing the mean ability ratings of children in different grades. We expected a positive relation between grade and ability. Fig. 5 shows the average ability rating for each grade and domain. As expected, children in older age groups had a higher rating than children in younger age groups. In all four domains, there is an overall significant effect of grade: addition $F(5, 1456) = 1091.4, p < .01, \omega^2 = .78$, subtraction $F(5, 1363) = 780.5, p < .01, \omega^2 = .74$, multiplication $F(5, 1215) = 409.6$, $p < .01, \omega^2 = .62$, and $F(5, 973) = 223.31, p < .01, \omega^2 = .53$ for division. Levene's tests show differences in variances for the domains

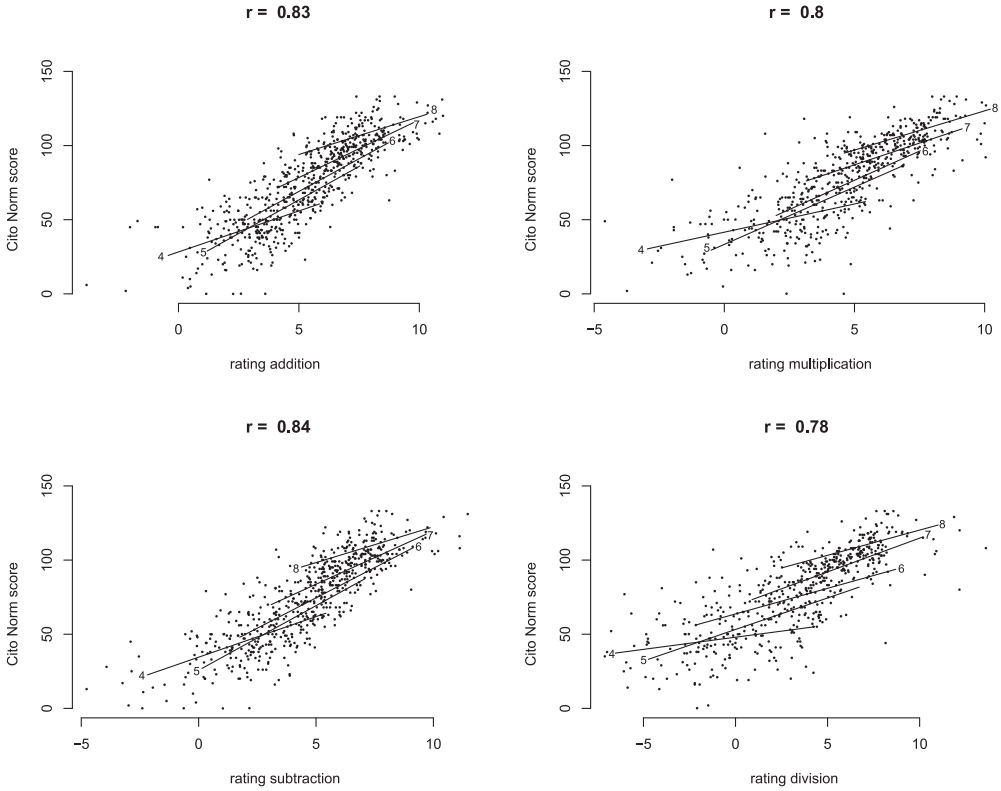

Fig. 4. Correlation between Maths Garden rating for the domains addition, subtraction, multiplication and division and the norm referenced CITO scores (mid 2008). Included are regression lines for each grade indicated by grade numbers.

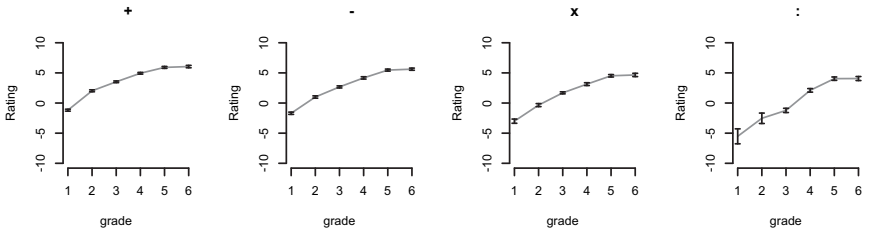

Fig. 5. Rating per domain and age group.

multiplication and division. However, the non-parametric Kruskal–Wallis tests also show significant differences for these domains: $\chi^2(5) = 753.28, p < .01$ for multiplication and $\chi^2(5) = 505.17, p < .01$ for division. For all domains, post hoc analyses show significant differences between all grades, except for the differences between grades five and six.

3.3. Reliability

One way to assess reliability is to compare children's ratings across domains. Since the domains involve many related operations we expect high correlations between them. The correlations between the four domains, addition, subtraction, multiplication and division, vary between .67 and .88, all significant at $p < .01$, indicating fairly high correlations. Another relatively simple way to assess the reliability of the Maths Garden is to construct parallel tests. We can compare item difficulty β's of so-called mirrored items (e.g. $7 + 4, 4 + 7$ and $2 \times 4, 4 \times 2$) for the domain of addition ($N = 48$) and multiplication ($N = 81$). Mirrored items should have very similar β's. Fig. 6 shows the correlation of the mirrored item β's for the domains addition and multiplication. These correlations are .88 and .98 ($p < .01$), respectively, indicating a high reliability of these item sets.

Besides difficulty β's, we can also compute the discriminatory power of items, which indicates how well the item discriminates low from high ability subjects. We estimated these so called a-parameters by using a logistic regression analysis on the accuracy responses predicted by the difference in rating between item and respondent. As in the preceding analysis, we compared the discriminatory power between mirrored items. The scatter plots in Fig. 7 show rather high significant ($p < .01$) positive correlations. The correlations for addition and multiplication are .74 and .71.

As a final test of reliability, we investigated the stability of the difficulty ratings β. A high correlation between β values of items at two time points far apart indicates high reliability. Therefore we would expect a stable item bank to correlate highly over time. We first looked at the correlation between the item β ratings, as they were set at the start of the project (week 36) and the item ratings in all subsequent weeks. In Fig. 8, this correlation is shown by the solid line. Clearly, the initial ratings, set on the basis of an analysis of Maths materials used by the schools, were quite good, as the correlation between initial ratings and the ratings after 40 weeks is still .85. Secondly, we also correlated established item ratings in week 44 with all item ratings in subsequent weeks (dotted line). This shows that these established ratings are very stable as the correlations in 32 weeks all stay above .95.

3.4. Item reuse

As a result of the longitudinal nature of the Maths Garden system, items are presented to the same child more than once. Although the system ensures that at least 20 other items are administered before an item is reused, this reuse may present a threat to the assumption of local independence (e.g. the response to an item must not depend on the previous response to the same item). To test this, we performed regression analysis with both the number of items and the amount of time between two presentations of the same item to the same child as predictors for the child's performance on that item. The child's performance was measured by subtracting his expected score $E(S_i)$ from the actual score S_i. If there is an item-specific learning effect, any child that encounters an item for the second time is likely to have a higher than expected score for that item. We selected pairs of data points that represented subsequent presentations of the same item to the same child. We selected the data so that no child contributed more than one pair of data points, resulting in $N = 478$ pairs of data points. Because

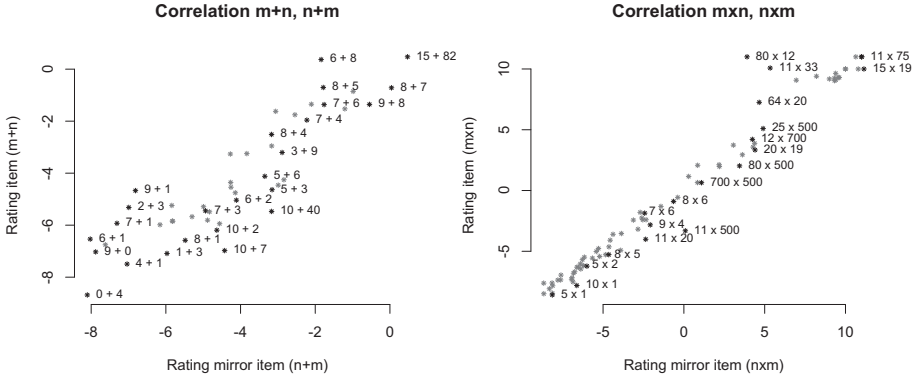

Fig. 6. Scatter plot of difficulty β's of mirrored items. Included are some examples indicated with black dots.

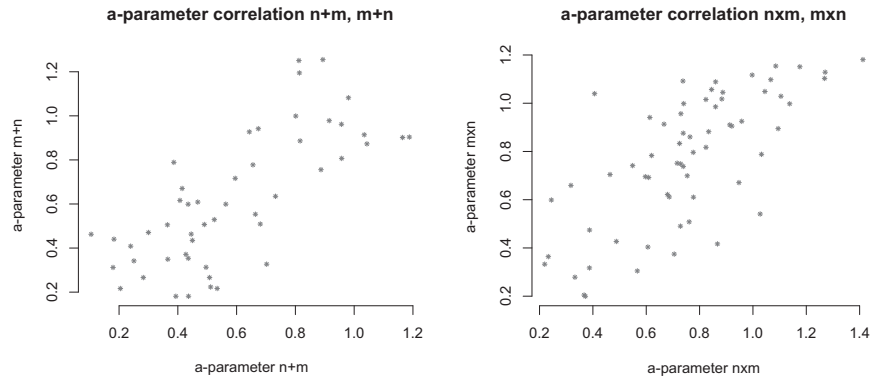

Fig. 7. Scatter plot of discriminatory *a*-parameters for mirrored items.

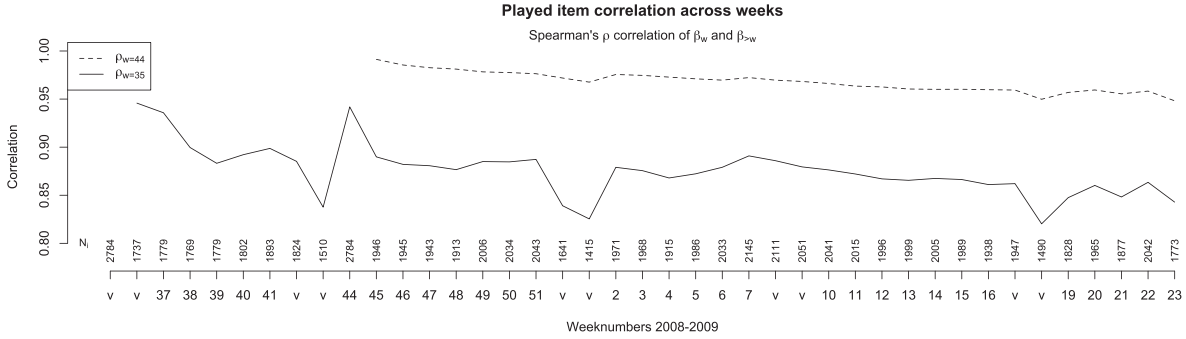

Fig. 8. Stability of items ratings for initial ratings (solid line) and established ratings after 2 months (dotted line). The *x*-axis displays week numbers (v = vacation). Correlations are computed over active (played) items in each week (N_i = amount of administered items).

item-specific learning effects are logically more likely to occur if there is a small amount of time between two presentations of the same item to the same child, we removed 90 data points with more than 30 minutes between the two presentations of the item. A regression analysis with this dataset shows no main effect for either the number of items, or the amount of time between two presentations of the same item to the same child: number of items, $R^2 < .001, F(1,476) = 0.39, p = .53$ and amount of time, $R^2 < .001, F(1,476) = 0.0072, p = .93$.

3.5. Maths Garden aims

In order to keep children motivated, items were sampled so that children solved about 75% of the items successfully. However, in the first few months we imposed a success rate of 70%. Fig. 9a shows the proportion of correctly answered items per grade and domain. Only the results of the children who answered more than fifteen items were included in the graph. The graphs show that the proportion of correctly answered items varied between .6 and .8 for most children. The proportion correct seems to be somewhat lower for subtraction and lower still for multiplication and division. At the start of this project, the domains addition and subtraction were briefly available for the lower age groups. This resulted in a lot of question mark use in these domains. To counter this unwanted effect we made the availability of these domains dependent on the proficiency on addition and subtraction. In total, the amount of question mark use in the Maths games was about 7.3%. Filtering out the question mark responses (Fig. 9b) results in considerably higher proportions correct.

One of the aims of the Maths Garden was that it should be a challenging web environment for children of all competency levels. The usage statistics can answer the question whether children are motivated to play the Maths games. They provide an indication of how attractive and challenging the children found the Maths games. It is possible that children visit the Maths Garden site mainly because their teachers told them to. To assess how intrinsically motivated the children were to play the games, we looked at the days and hours that children played in the Maths Garden. Fig. 10 (top) shows the number of solved arithmetic problems per day of the week and Fig. 10 (bottom) shows the number of solved items per hour of the day. Not surprisingly, most problems were solved on Monday till Friday and between 9.00 a.m. and 3.00 p.m. However, both graphs also show that a considerable number of problems were solved after school hours and during the weekends. Actually, 33.2% of all problems were solved outside school hours.

To investigate whether competency had any effect on the motivation, we looked at the relation between ability and playing frequency. Only data of children who solved 15 or more problems was included to ensure accuracy of the ability estimates. We found only low but significant ($p < .01$) correlations between ability level and playing frequency for all domains. The correlations for the domains addition, subtraction, multiplication and division were, −0.15, −0.12, −0.05, and .09, respectively. The playing frequency does not appear to depend importantly on the competency level of the children.

3.6. Diagnostic ability

We will briefly demonstrate the diagnostic and tracking ability of the Maths Garden by considering a few examples. Using the high frequency dataset, we were able to provide individual and group diagnostics. Fig. 11 shows the percentage of typical errors a given child had

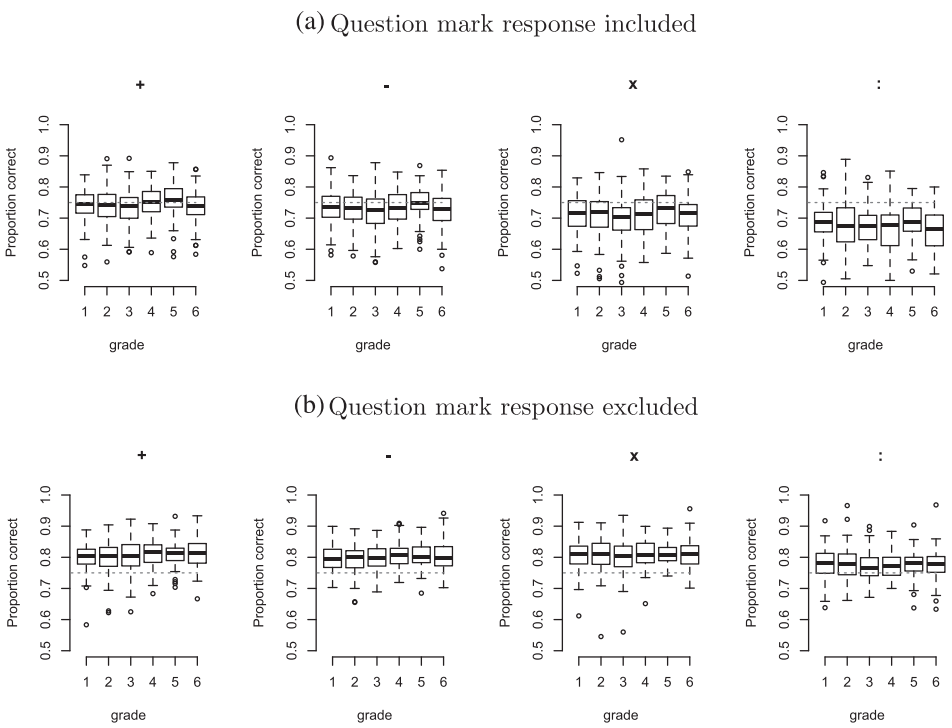

Fig. 9. Proportion correct per grade and domain.

made (bars) compared to the percentage of these errors made by children of the same grade (solid line). We can see, for instance, that this child makes significantly more zero errors (400−200 = 380), for the domain subtraction than other children in the same grade. We provided teachers with such graphs for individuals and groups of individuals (e.g. for the whole class).

Detailed analysis of the item difficulties provides us with insight into sources of item difficulty. Some interesting results have emerged. For example multiplications by 10 or even 100 and one digit numbers (7×100) are among the 10% easiest items for this domain. In division it appears that items of the type *nn*:*n* (77:7) are also very easy (again among the 10% easiest items). Straatemeier et al. (submitted) tested how well all kinds of item effects, previously studied in isolation, predict item difficulty. The combined item effects, such as problem size, ties and the 5 effect, explained 90% of the variance in the difficulty of simple multiplications items.

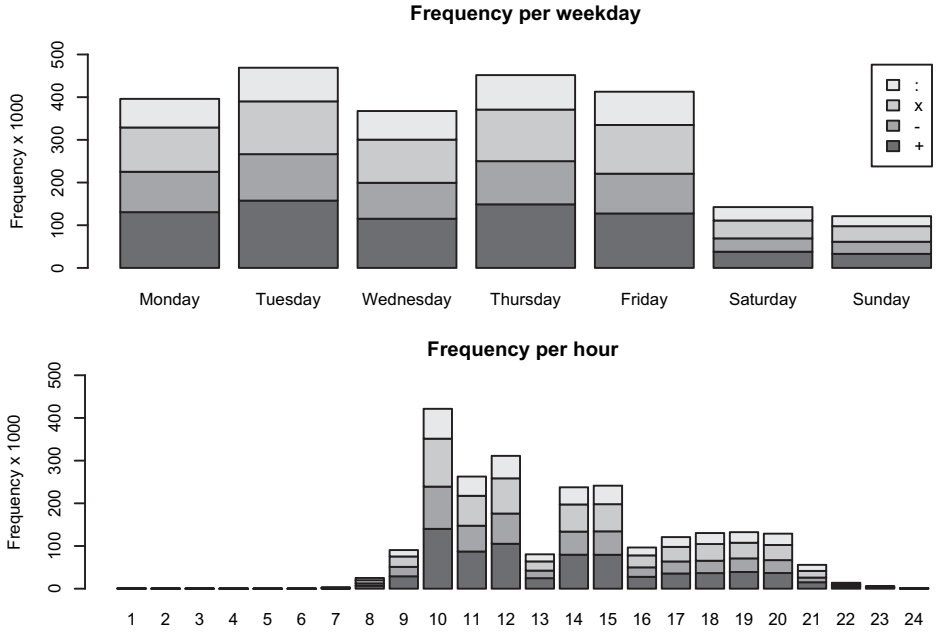

Fig. 10. Playing frequency during the week and during the day.

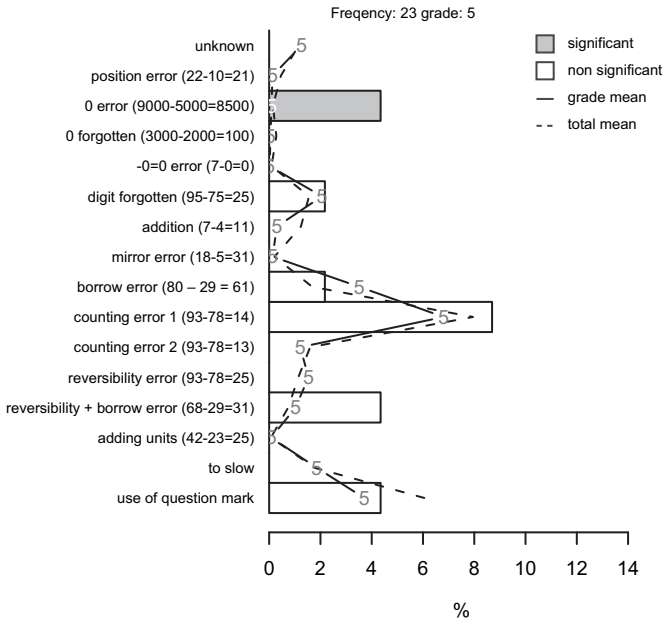

Fig. 11. Error analysis of answers to subtraction problems of a child in grade 5. Bars display the percentage of errors for this child in a specific week. The lines display the percentage of errors made by other children in the Maths Garden (dotted line) and by other children in grade 5 (solid line).

3.7. A window on developmental change

The high frequency measurements combined with the size of the sample, provide unique insights into arithmetic development and learning trajectories of children. In the Maths Garden, trend analyses are provided to teachers. Fig. 12 shows the progress of a single child compared to all other children in the same age group. Teachers can use this information to consider interventions. As can be seen in the graph, this child started out having an average rating and a flat growth curve. By week 45 this child started to acquire the necessary ability and by week 49 the child was in the top 25% of all children.

At micro level it is even possible to study the learning pattern of one child on a specific item over time. For example, in graph 13 we see the answers and response times of two children on two items across weeks. In the top graph of Fig. 13 we see an individual who did not know the answer to the Maths question 9×9 and answered with a question mark in about 5 to 10 seconds at the first ten occasions. Then there were two mistakes where the child joined the two digits instead of multiplying. However, in the next attempt the question was answered correctly but more time was needed to respond. From this point on the ability level seems sufficient for consistent correct and speedier answers. The bottom graph of Fig. 13 shows a lucky guess in the first week (third trial) followed by a gradual gain in insight. Half way week 42 this child started answering correctly more often but with highly varying response times. At the end of week 44 the response time droped. Note that occasionally errors keep occurring. These examples illustrate the level of detail that is possible in the analysis of Maths Garden data.

4. Discussion

In this paper we presented and tested a new model for computerized adaptive practice and monitoring. The results concerning the validity and reliability are promising. The high correlations with the norm referenced CITO scores indicate high criterion validity. The increase in player ability rating across grades also supports this, although the children in grades 5 and 6 did not seem to differ. This is probably due to the fact that in the domains we tested no new mental arithmetic techniques are taught in grade 6.

By simulation, we compared measurement precision and measurement bias of CAP to standard CAT. For easy items the use of the HSHS scoring model, which combines speed and accuracy and the Elo rating system (ERS) resulted in less loss in measurement precision and less

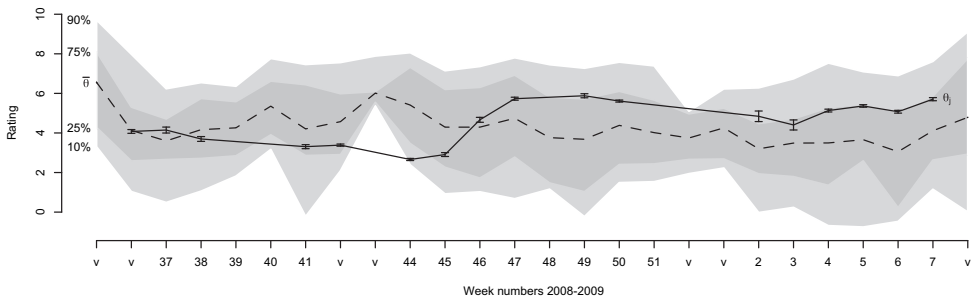

Fig. 12. Progress chart of a child in grade 6 (black line), in comparison to the mean of grade 6 (dotted line).

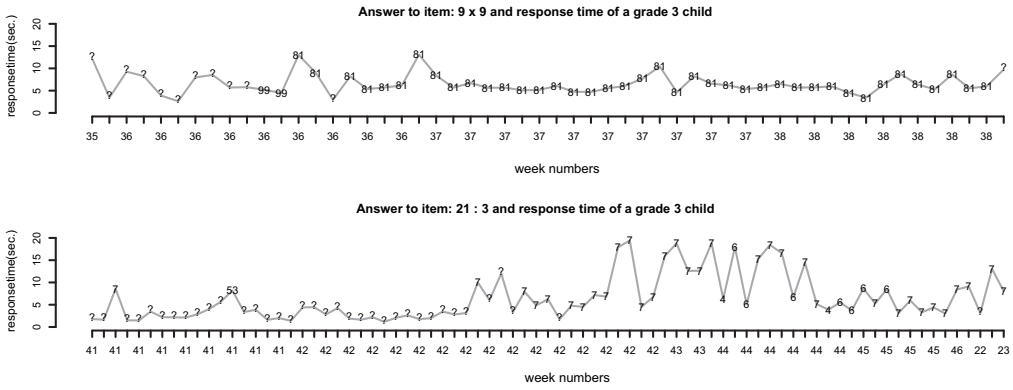

Fig. 13. Response time pattern for two children on different items during a number of weeks (*x*-axis). The *y*-axis indicates the response time in seconds. The answer is displayed in the graph. The question mark answer means that the child pressed the '?' button.

bias than found in standard CAT estimation. The ERS combined with the 1PL model, using only accuracy data, resulted in worse estimations. Concerning the items and the item bank, we found that difficulty ratings converge in about eight playing weeks, resulting in consistent difficulty ratings across time. High reliability is also indicated by the high correlations of the difficulty and discrimination parameters between sets of mirrored items. We have not found any indication of learning effects caused by the reuse of items, therefore also indicating the assumption of local independence has been met for reuse of items. However in other learning domains this issue still requires careful consideration.

The fit statistics for the HSHS model are still in development, and are therefore not included in the result section of this paper. Evaluation of the goodness of fit for IRT models is an active area of research, and so far definite solutions are lacking (Embretson & Reise, 2000). Some of the relevant issues (Hambleton, Swaminathan, & Rogers, 1991) concern, the sensitivity of the chi-square fit statistic to sample sizes, technical issues in the testing of dimensionality (Hattie, 1984, 1985), and the testing of the assumption of local independence. Evaluating the fit of IRT models is more complicated still in the context of computer adaptive testing, due to the inherent incomplete item-person data matrix. An alternative approach to comprehensive model fitting consists of checking model assumptions, and establishing reliability and validity (Hambleton & Swaminathan, 1985). Here we have sufficed with this alternative approach.

We can conclude that children were motivated to play the Maths games. The frequency data demonstrated that children played a lot outside of school hours. Children with a lower ability in Maths did not play appreciably less, which suggests that they found the Maths games as motivating as high ability children did. We demonstrated that the Maths Garden has many possibilities as a diagnostic tool. The error analysis can provide teachers with valuable insight into the kind of errors that individual pupils make. This information can be used to optimize interventions. The current dataset, consisting of a large number of individual high frequent time series, allows for many further investigations of difficulty effects (Straatemeier et al., submitted for publication), strategy patterns in mathematical problem solving and individual learning trajectories. The item ratings also provide insight into what we call informal learning paths. Because of the adaptive item ratings, we gain an on the fly insight into the difficulty of arithmetic problems. Some items turned out to be unexpectedly easy. For instance, $8 + 6$, $5000 + 5$ and $50 + 60$ were almost equally difficult whereas $8 + 6$ is taught much earlier on in the Dutch curriculum than the other two addition problems. This kind of information can be used to determine the curriculum (i.e. what is taught) in each grade.

One of the problems with the Elo rating system is the occurrence of rating inflation and deflation (Glickman, 1999), which we call drift. In educational applications, one source of drift is that new young players start with low ratings and stop playing when they leave school with high ratings. This causes a systematic downwards drift in item rating and, as a consequence, lowers person ratings. This does not seem to jeopardize the operation of the Maths Garden, since drift influences player and item ratings simultaneously. The main problem lays in the interpretation of the rating. Rating points cannot be accurately compared following inflation or deflation. Therefore we present transformed ratings to teachers and users to prevent interpretation problems. Transformation is conducted by calculating the average probability correct for a single user on all items in the domain, as shown in equation 10:

$$\overline{P} = 1/n \sum_{i=1}^{n} \frac{1}{1 + e^{-a\left(\widehat{\theta}_j - \beta_i\right)}}. \tag{10}$$

This value is an estimation of the percentage of items in the domain that the user is able to answer correctly. We also reduced drift by incorporating the rating uncertainty in calculating the *K* factor, which minimizes the influence of unreliable person and item estimations on the updating proces. A related issue is the convergence speed. This is the time or number of responses needed to get a stable rating. We set the rating uncertainty parameters of the *K* factor, which determine the convergence speed, on the basis of extended testing. A better approach would perhaps be to estimate the uncertainty based on aberrant response patterns, where unexpected responses are used as an indication of unreliability.

A last issue concerns the one-dimensionality of the Maths domains. In practice, every test and item bank is expected to violate the assumption of one-dimensionality to some degree. Though we see no immediate effects on ability estimation the question of how robust the HSHS Elo algorithm is to violation of this assumption needs further investigation. We also intend to further address the possible individual differences between children and how the HSHS scoring rule affects their behavior.

In conclusion, Maths Garden meets the requirements we set for the practice and progress monitoring system. It is worth noting that although the new CAP algorithm is implemented in the domain of Maths, the system can be applied to all kinds of learning domains. In the

new release of Maths Garden more games, e.g. fractions, have been added and a language garden is in development. Also, the number of schools using the Maths Garden continues to grow steadily (about 150 in October 2010), yielding about 50 thousand responses per day. We expect a fast adoption of computers, such as handhelds, minicomputers and tablets, in primary schools in the next 5 years. If children do their daily exercises in practice and progress monitoring systems using these devices, we expect many benefits for students, teachers and scientists.

References

Batchelder, W. H., & Bershad, N. J. (1979). The statistical analysis of a thurstonian model for rating chess players. *Journal of Mathematical Psychology, 19*, 39–60.

Commissie Dijsselbloem. (2008). *Parlementair onderzoek onderwijsvernieuwingen*. In *[Parliamentary inquiry educational reform]*, . The Hague: Sdu publishers.

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*, 379–393.

Elo, A. (1978). *The rating of Chessplayers, Past and present*. New York: Arco Publishers.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists* (1st ed.). Mahwah: Lawrence Erlbaum.

Ericsson, K. A. (2006). *The cambridge handbook of expertise and expert performance*. chapter The Influence of experience and Deliberate Practice on the Development of Superior Expert Performance. Cambridge University Press.

Expertgroep doorlopende leerlijnen. (2008). *Over de drempels met taal en rekenen*. In *[crossing the barriers of language and mathematics]*, . Nationaal expertisecentrum leerplanontwikkeling.

Fullan, M. (2006). The future of educational change: system thinkers in action. *Journal of Educational Change, 7*, 113–122.

Glickman, M. (1995). A comprehensive guide to chess ratings. *American Chess Journal, 3*, 59-–102.

Glickman, M. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics, 48*, 377–394.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory:principles and applications/Ronald K. Hambleton, Hariharan Swaminathan*. Distributors for. North America, Kluwer Boston, Boston:Hingham, MA, U.S.A: Kluwer-Nijhoff Pub.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory/Ronald K. Hambleton, H. Swaminathan, H. Jane Rogers*. Newbury Park, Calif: Sage Publications.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49–78.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.

Janssen, J., & Engelen, R. (2002). *Verantwoording van de toetsen Rekenen-Wiskunde 2002*. In *[Justification of the Cito student monitoring system for arithmetic performance, 2002]*, . Arnhem: Citogroep.

Kraemer, J., Janssen, J., Van der Schoot, F., & Hemker, B. (2005). *Balans van het reken-wiskunde onderwijs halverwege de basisschool 4[Fourth assessment of the mathematics education halfway primary school]*. Arnhem, The Netherlands: CITO.

Maris, G. & Van der Maas, H. (Submitted for publication). Scoring rules based on response time and accuracy.

Van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308.

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

Van der Linden, W., & Hambleton, R. K. (Eds.). (1997). *Handbook of Modern item response theory* (1st ed).. Springer.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores; with contributions by Allan Birnbaum*. Reading, MA: Addison-Wesley.

Van der Maas, H., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology, 118*, 29–60.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmark Paedagogisk Institut.

Ratcliff, R., & Rouder, J. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*, 347–356.

Straatemeier, M., Jansen, B., Klinkenberg, S., & Van der Maas, H. (submitted for publication). *A large-scale integrated analysis of basic multiplication effects, using a new computerized adaptive progress monitoring system*.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: a DMAT primer. *Behavior Research Methods, 40*, 61–72.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Mislevy, R., et al. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.