Math Garden: A new educational and scientific instrument

Straatemeier, M.

Citation for published version (APA):
CHAPTER 7

Summary and general discussion
The aim of this thesis was to study the development of mathematical ability. Early in the research project we concluded that it was necessary to construct a new measurement tool in order to study our questions concerning the dynamics of development of mathematical knowledge and abilities. The main focus of this thesis therefore shifted to the construction of a new measurement device: Math Garden. We aimed to establish the working and value of Math Garden as a measurement tool, which is a prerequisite for further research using Math Garden. This thesis is, therefore, only the starting point of the research that can be performed with Math Garden. In this final chapter we will answer the question whether Math Garden is indeed a valuable measurement instrument. We will summarize the findings of this thesis, obtained with Math Garden, but also the challenges and potential problems faced when using this method. We especially pay attention to the value of response times for the assessment of math abilities. The integration of response times and accuracy is often considered to be problematic in psychological assessment and therefore deserves further discussion. We consider whether Math Garden meets the goals set out in the Introduction and discuss further possibilities for extensions of the system. Finally, we list ongoing research projects and scientific challenges for the future.

Summary of main findings and future challenges

Computer adaptive measurement of development

In Chapter 1 we argued that a microgenetic design (i.e., high frequent measurements) is necessary to study the complex dynamics of mathematical learning and development. We decided to construct a measurement tool, Math Garden, that both children and teachers were motivated to use in a high frequent manner. We concluded that one of the main requirements of Math Garden was that it should utilize the technique of computer adaptive testing (CAT) to measure children’s mathematical ability. With CAT a person’s ability is measured dynamically: item administration depends on the person’s previous responses, thereby tailoring the test to the person’s ability. CAT, therefore, enables precise measurement of abilities of a wider range and with fewer items than standard fixed tests (van der Linden & Glas, 2000; Wainer, 2000). CAT can save a lot of time when repeated measurements are required. In addition, we assumed that CAT would have a positive effect on children’s motivation because they rarely make items that are too easy or too difficult, thereby possibly stimulating high frequent use of Math Garden.

In Chapter 2 we argued that standard CAT is not suited for use in computer adaptive practice systems such as Math Garden. We developed a new CAT method with which we aimed to solve three problems concerning the standard CAT technique, namely that
1) items need to be pre-calibrated in standard CAT, 2) standard CAT doesn’t incorporate response times, and 3) standard CAT operates most effectively when the probability of success on administered items is .5. These are serious problems. Pre-calibration is expensive and time-consuming. Ignoring response times is not only a waste of data but may also hide strategic differences between subjects in their choice of the speed-accuracy trade-off. Finally, a probability of success of only .5 is demotivating for most children (and adults).

We developed a novel extended CAT approach based on the Elo rating system (Elo, 1978) in which both item difficulty (item ratings) and person ability (person ratings) are updated after each answered item. Within this system items are calibrated on the fly making it unnecessary to pre-calibrate items. In addition, we use the new High Speed High Stakes (HSHS) scoring rule of Maris and Van der Maas (2012) to incorporate response times into our CAT method. With this scoring rule both punishment for incorrect responses and reward for correct responses decrease linearly with response time. Eggen and Verschoor (2006) demonstrated that with standard CAT measurement bias increased and measurement precision decreased considerably when easy items were selected for administration. In Chapter 2 we demonstrated with simulation studies that the standard CAT indeed performs best when administering items with a probability of success of .5, but that the Elo system combined with the HSHS scoring rule resulted in less bias and higher measurement precision for easy items compared to standard CAT. Levels of bias and measurement precision were acceptable with the extended CAT approach. The use of response times is essential in this approach, as the Elo system using accuracy alone performed worse than standard CAT. Thus, the extended CAT approach enables the administration of easy items while still effectively measuring ability.

The results concerning the reliability and validity of the ratings resulting from the extended CAT approach, as reported in Chapter 2, are promising as well. The item ratings converged to stable item ratings in about eight weeks of playing in an active participant sample of 3648 users. So, with the extended CAT approach we were able to obtain calibrated items within a relatively short period of time. With the current number of active users (> 85,000, September 2013) stable item ratings are reached in an even much shorter time period. This makes Math Garden a platform for the fast development of new adaptive games. When a new item bank is available, it can be implemented in a Math Garden game and the item bank is calibrated within a few days. Other indications of high reliability of the item ratings are the high correlations of the item ratings and discrimination parameters between sets of parallel items.

To assess the reliability of person ratings, it is not possible to use the stability of the ratings over time. We expect children to improve in math over time and, therefore, also expect
person ratings to vary over time. We can, however, compare person ratings across domains. We found fairly high correlations (range: .67 - .88) between the domains addition, subtraction, multiplication, and division. These high correlations are also an indication of the validity of the person ratings. Other results supporting the assumption that the extended CAT approach yields valid person ratings are the high correlations between person ratings on these four domains and the general math ability scale of the pupil monitoring system of Cito (Janssen & Engelen, 2002) and the positive relation between grade and person ratings. The issue of validity was further addressed in Chapter 3 and 4 and we will address the implications of these studies below.

Summarized, in Chapter 2 we demonstrated promising results concerning the reliability and validity of the extended CAT approach. With this new approach items do not need to be pre-calibrated and the approach also works well when easy items are administered. The use of response times is an essential and necessary feature to achieve this. However, there are still several challenges when using the Elo system and several issues need to be addressed.

The first issue, drift, is a fundamental problem of Elo type rating systems and of the general class of paired comparison systems to which the Elo system belongs (Glickman, 1999). The well-known IRT models also belong to this class. What counts in all these models are the differences between Elo ratings of persons, or between item and person ratings. All ratings are measured at the same scale, which has an arbitrary zero point. In IRT the zero point is normally set to a fixed value such as the difficulty rating of a certain item or the mean of the ability ratings. In Elo systems this is impossible as the ratings are updated continuously. We developed and tested different solutions for drift, some of which are discussed in Chapter 2.

The second issue, unidimensionality, is also well known in IRT. Our item banks contain items of a wide range of difficulties. Do all these items measure the same trait? Is 23.3 x 6.5 more difficult than 3 x 6 because a higher level of multiplication skill is required to answer the item correctly or does the item also measure other math abilities? A new theoretical approach to this issue can be found in Van der Maas, Molenaar, Maris, Kievit, and Borsboom (2011), who propose a new IRT model for abilities derived from a mathematical process model (the Ratcliff diffusion model). This Q-diffusion IRT model has several remarkable properties such as a true zero point for the ability scale. It also implies a new definition of unidimensionality. In the Q diffusion IRT model any person with positive ability, even very small, will solve all items in a test given sufficient time. As an example, van der Maas et al. (2011) use the ability to move. Any person with the ability to move will pass all items (all distances) given enough time. This very strict definition of unidimensionality clearly poses problems for our Math Garden games. Multiplication items 3 x 6 and 23.3 x 6.5 would then
probably not be acceptable in one test. Knowing decimal numbers is an additional ability not required for solving 3 x 6. It is an open question whether we can develop CAT methods based on the Q-diffusion model.

In practice, however, we do not expect unidimensionality problems with these items or the multiplication game in general as the abilities that are necessary to solve the items will co-vary. Moreover, we are rather strict with respect to the requirement of unidimensionality in Math Garden. Separate rating systems are used for the different math games; measuring traits separately that are often measured together in other math tests, such as the widely adopted Cito tests. The practical approach to unidimensionality in IRT is to test for unidimensionality statistically. As long as the traits of knowing to multiply and knowing decimal numbers co-vary in the population under study, dimensionality tests will indicate unidimensionality. A remaining challenge is to develop unidimensionality tests for adaptive testing systems. The Math Garden data set does provide new possibilities to do this as some skills, such as addition, subtraction, multiplication, and division, are both measured in separate games for each operation and in mix games.

It should be noted that our practical defense of unidimensionality, based on the co-variation of various math abilities in the population, matches educational practices in schools. Most math methods teach math skills in the same order. If, however, various schools would adopt different teaching methods in which math skills are taught in a very different order, our games could start to suffer from multidimensionality. But again, Math Garden provides a platform to test for these effects.

Initially Math Garden contained a balance scale game (Hofman, Visser, Jansen, & van der Maas, submitted), inspired on the Piagetian balance scale task for proportional reasoning. This task is known for its violations of unidimensionality, as performance across age groups is characterized by the use of qualitatively different problem solving strategies, associated with different rankings of item difficulties (e.g., Jansen & van der Maas, 1997, 2002). Indeed, certain items were solved correctly by young children but not older children and this led to unstable item ratings. We decided, therefore, to remove the balance scale game from Math Garden. This example illustrates that large violations of unidimensionality can indeed be a problem for Math Garden but also that we were able to detect these violations.

The last issue concerns standard errors of ability ratings in the Elo system. Standard errors are very important when reporting ratings and rating developments in order to evaluate the reliability of ability estimates. However, it is not easy to develop a sound statistical method for computing standard errors for Elo ratings. Currently, Matthieu Brinkhuis and Gunter Maris develop new adaptive rating systems that do allow the calculation of standard errors for ratings.
What affects problem difficulty?

In Chapter 2 we investigated the validity of the rating system from a person perspective by correlating the person ratings of several math games with each other and with performance on independent math tests. In Chapters 3 and 4 we investigated the validity of the rating system from an item perspective. If the item ratings resulting from the extended CAT system are a valid measure of item difficulty, we should be able to explain variance in item difficulty by problem characteristics found to predict item difficulty in previous research.

Chapter 3 and 4 are similar in structure. First, we identified problem characteristics reported in the literature as affecting item difficulty. Second, we included all problem characteristics in an integrated analysis. This is in contrast to most studies, in which these effects are mostly studied in isolation. Simultaneous analysis of all known problem characteristics enabled us to detect which problem characteristics independently affected problem difficulty, that is, when other effects are accounted for. In addition, the microgenetic setup of Math Garden enabled us to study the robustness of effects over time by performing replication studies over time. Problem characteristics that are assumed to have a substantial effect on item difficulty should affect item difficulty at each time point.

In Chapter 3 we investigated the item difficulties of simple multiplication problems, that is, problems ranging from $1 \times 1$ to $9 \times 9$. We were able to predict the difficulties of these problems very precisely. The explained variance in item difficulty equaled 88% for a test format with open-ended items and 90% for a test format with forced-choice items. In addition, these models included many of the effects found in previous research. Problem size increased item difficulty, whereas the inclusion of the numbers one, two, five, and nine decreased item difficulty. These effects were robustly affecting item difficulty across months. In addition we found that the inclusion of these special numbers interacted significantly with problem size: The effects were stronger for problems with large numbers than for problems with small numbers. The presence of a tie (two similar operands) also decreased item difficulty, but the results concerning this characteristic were mixed. When using open-ended items, an interaction with problem size was found. When using multiple-choice items, only the main effect was found.

In Chapter 4 we investigated the item difficulties of both simple and complex addition and subtraction problems. We were able to explain a large proportion of variance in item difficulty of simple and complex addition problems (both 91%) and simple (87%) and complex (86%) subtraction problems. Again we found a significant positive contribution of problem size to item difficulty. The best operationalization of problem size differed between the selected item sets. For example, for complex problems the number of digits in the problem (assumed to be an indication of the number of sub results that need to be
calculated) had a strong increasing effect on item difficulty. For both simple and complex addition and subtraction problems the problem characteristic that had the largest and most robust increasing effect on item difficulty was whether carrying (addition) or borrowing (subtraction) was required to solve the problem. Again the presence of particular numbers in either the problem or the answer was found to affect item difficulty. Because we also studied complex problems we were able to study the effect of the presence of decades in the problem, which significantly decreased item difficulty. This finding suggests that taking knowledge of the base-ten system helps when solving complex problems. Overall the models for addition and subtraction problems were very comparable. An important difference was the order effect. There was a substantial order effect for both simple and complex subtraction problems. Overall subtraction problems in which the subtrahend is smaller than the remainder (e.g., 16-7=9) are easier than their mirrored problems in which the remainder is smaller than the subtrahend (e.g., 16-9=7). No order effect was found in the item ratings of addition problems.

The results presented in Chapter 3 and 4 support the validity of the item ratings. Large proportions of variance in the item ratings of addition, subtraction, and multiplication problems were explained by a limited set of problem characteristics, and the estimates were robust over time. Moreover, the majority of these effects could be explained by results reported in the literature. In previous studies either error rates or response times where used as an indicator of item difficulty. We showed that similar effects could be found in the item ratings, in which accuracy and response times were integrated. In conclusion, the extended CAT approach used in Math Garden leads to stable and well interpretable item ratings, supporting the validity of the system.

What do these item ratings tell us about children’s mathematical development? The item ratings are based on the answers of all children using Math Garden, but children do not answer each item, due to the CAT approach. They only answer items that are developmentally relevant for them, that is, suited for their ability level. We therefore expect that the item ratings, obtained in a sample that includes children from all grades of primary school, reflect the whole learning process during primary education. Ordering the items by item difficulty gives us insight into the order in which children in primary education master math problems. Selecting seven addition problems from the addition item set with an answer below 100 and ordering these by item rating leads to the following sequence: 2 + 3, 80 + 10, 77 + 1, 3 + 21, 7 + 8, 22 + 38, and 28 + 34. Figure 1 shows the development of the item ratings of these items, showing that this order was consistent across months, despite fluctuations in the estimates. Primary school teachers would immediately see that this order does not match the order in which these problems are taught in most math meth-
ods. For example, $7 + 6$ was found to be more difficult than the problem $80 + 10$. In most math methods children first learn all problems with answers up to 20 before being asked to solve problems with larger operands. In Chapter 4 we concluded that carrying is the problem characteristic that has the largest increasing effect on item difficulty, which explains why $7 + 8$ is more difficult than the problems $80 + 10$, $77 + 1$, and $3 + 21$. Insight into the problem characteristics affecting item difficulty and the interactions and dependencies between these characteristics can provide us more information about the challenges children face when mastering mathematical problems. Using this information in the development of math methods may improve education.

![Figure 1](image_url)  
Figure 1 Development of item ratings of seven addition problems from 1-1-2012 till 1-9-2013.
There are, however, also drawbacks when using Math Garden data for studying development in mathematical abilities. As children answer only a subset of items, it is impossible to draw conclusions on how children of different age groups perform on the same item set. For example, most children in the higher grades in primary education will rarely answer the simple multiplication problems studied in Chapter 3, as their probability of answering these items correctly is very high and these items are therefore not presented (frequently) by the extended CAT method. Based on the assumptions that 1) items that are not presented are too easy (when item rating is far below the person rating) or too difficult (when item rating is far beyond the person rating); and 2) the rating scale is unidimensional, we can compare a large set of problems with a broad range in item difficulty without the necessity of administering the same problem set to participants of various ages.

It is not so easy to convince developmental researchers of the value of this type of data as standard age comparisons and certain types of individual analyses are not possible with the data set. In addition, item difficulty has rarely been used as a dependent variable in developmental research. Besides demonstrating the value of item difficulties as a measure for developmental research, as we have done in Chapters 3 and 4, we also see other possibilities with the Math Garden data set to meet with these hesitations in the field. We could, for example, recalculate the item ratings for only a subset of children from the sample. This can be achieved by using their logged records of response times and accuracy on attempted problems and then rerunning the estimation procedure. Recalculating item ratings for different age groups makes it possible to study age differences. We have, for instance, followed this approach in a study concerning children’s skills to enumerate visual displays of one to six elements (Jansen et al., 2014). Recalculations of item ratings per age group is only advisable when the recalculation sample has answered a sufficient number of the problems under study and when children in the recalculation sample are a good representation of the age group. That is, not only weak or proficient children should be included.

To study individual differences, one could also look at the difference between a child’s expected score on an item and her/his actual score. Given a child’s rating and the rating of the presented item, an expected score is calculated. Deviations from the expected score can be considered indications of deviations from the general developmental curve. For example, if a child consistently has a lower score on certain items compared to what was expected, one could conclude that these specific problems are more difficult for this individual child than for children of comparable mathematical ability.
Classification of children's errors

In Chapter 5 we used another source of information of the Math Garden data set to study children’s development of mathematical abilities, namely the errors they make. The data of Math Garden are particularly interesting for questions regarding development of mathematical strategies, as children solve problems that match their ability level and each child therefore makes a comparable number of errors. Errors may provide interesting information concerning the (possible erroneous) strategies children use. Error analyses can be performed at the individual level.

One difficult issue that arises when investigating children’s errors is how errors should be classified, as different erroneous strategies may result in the same error. For example, we could classify the answer 18 to the problem 9 x 9 into at least two error categories. One could argue that the child added both numbers instead of multiplying them. Another possibility is that the child has calculated the correct answer but reversed the numbers in his final answer. Other applicable error categories can be found in Chapter 5. This example is not an exception. In Chapter 5 we demonstrated that there was a large overlap between well-known error categories for multiplication problems, emphasizing the necessity for classification methods that deal with double classifications. The standard solution in the literature is to determine an order in which categories are treated. However, we don’t see any rationale for these orders.

In Chapter 5 we compared six different methods for classifying children’s errors and argued that the best method for classifying children’s answers is through use of the weighted frequency rule. With the weighted frequency rule the classification order is based on the frequency of error categories in the dataset. High frequent errors precede low frequent errors. The assumption is that the frequency is an indication of the plausibility of the error classification. However, to ensure that small categories and subcategories are not being overlooked the frequencies should be corrected for the number of possible manifestations of an error category. For example, an answer is classified as a ten table error if the answer is the correct answer to any other single digit multiplication problem. There are however 36 potential ten table answers. If no correction is used this category is always preferred over the error category “addition” (addition instead of multiplying), which has only one possible manifestation per problem.

The weighted frequency rule does not suffer from “literature bias”, which is a problem for literature based classification methods. In these methods definitions of error categories and the classification order are based on errors reported in previous studies. This prevents finding new error categories and leads to overestimating error categories that are high in the classification order. The weighted frequency rule is more data driven and thereby less
dependent on the decisions of the researchers. We analyzed children’s multiplication errors with the weighted frequency rule and found that the proportion of unclassified errors was lowest using this rule. We also concluded that the majority of children’s errors are the correct answer to a related multiplication problem, especially for simple multiplication problems (e.g., the answer 30 to the problem 7 x 5).

The present method is a large improvement to the methods used in research concerning the classification of arithmetical errors so far. Moreover, it is a method that has practical value for education, as it can easily be used for error analyses with a diagnostic purpose. A drawback of the weighted frequency rule is, however, the lack of statistical fit measures that enable conclusions about how well the classification model fits the data. Although methods usually used for this purpose (latent variable methods) are difficult to apply because of the complexity of the data, that is, the number of potential answers and error categories, we think that model based approaches for error classification should be developed in the future. The weighted frequency rule presented in Chapter 5 is a good starting point.

One could argue that it is difficult to disentangle age and item effects when investigating children’s errors with the Math Garden dataset. Children of different ages will answer different problems in Math Garden and this will lead to different types of errors. We believe, however, that this can also be seen as an important advantage of the Math Garden dataset. Because children answer items that are developmentally relevant for them, these errors reflect errors made in the classroom.

Response times and mathematical ability

In this thesis response times play an important role. We believe that the use of response times has several advantages above using accuracy alone. First, response time is related to arithmetic ability. More advanced strategies take less time than less advanced strategies (Lemaire & Siegler, 1995; Steel & Funnell, 2001). For example, using the count all strategy to solve the problem 7 + 3, which requires counting all fingers, will take more time than using the min strategy, which only requires counting on from the number 7 (“7, 8, 9, 10!”). Including both response times and accuracy may, therefore, improve the measurement of arithmetical abilities considerably.

Response times also play a role in educational assessment. First, scores on mathematics tests depend on response times. On speeded tests, such as “Tempotoets Automatiseren” (TTA; de Vos, 2010), children are explicitly instructed to solve as many problems as possible within a given period of time. On other tests, like exams, the role of response times is more implicit as a time limit is set for solving a set of problems. However, the number
of problems solved correctly defines the score on the test and individual differences in speed-accuracy trade-off are not taken into account. Response times also play a role in learning goals, although the goals with respect to response times are often formulated vaguely. The Dutch learning goals (Meijerink et al., 2009) include statements concerning the type of problems children should be able to solve “fluently” (NL: “vlot”), but no explicit statements are made concerning the maximum response time that is considered fluent. Because response time is related to arithmetic ability, we believe that it should have a more prominent (i.e., explicit) role in both the formulation of learning goals and in calculating achievement scores on math tests.

The inclusion of response times to measure arithmetical ability is especially relevant for easy items. As subjects tend to answer easy items correctly, accuracy on these items provides little discriminatory information and response times become an important extra source of information. For example, Van der Maas and Wagenmakers (2005) show that there is a negative relation between response time and chess ability and that this relation is especially strong for easy items. Indeed, results reported in Chapter 2 show that the extension of the Elo system with the HSHS scoring rule improved measurement precision and decreased bias for easy items compared to standard CAT estimations and the Elo rating system using accuracy alone. Hence by using the additional information of response times we were able to administer easy items in a computer adaptive testing procedure (probability of answering correctly equal to .75) while still attaining reliable ability estimates.

The use of response times in research concerning mathematical ability is very common but integration of response times and accuracy data is complicated by the trade-off between speed and accuracy. Large individual differences in how subjects value speed and accuracy have been found (Wickelgren, 1977). In addition, the instruction used when administering a test may influence these values. For example, Kirk and Ashcraft (2001) showed that specific test instructions either increased or decreased the use of the fast retrieval strategy. Without explicit instruction subjects are free in how they balance speed and accuracy, possibly leading to non-optimal trade-off decisions for some subjects. In Math Garden we used the HSHS scoring rule to solve the speed-accuracy trade-off problem. Given this scoring rule, an optimal trade-off between speed and accuracy exists and this optimal trade-off is similar for all subjects.

An important characteristic of this scoring rule is that punishment is highest for fast responses, which may be the result of guessing. Hence, the HSHS scoring rule encourages deliberate and thoughtful responses. At first it may seem odd that a fast incorrect response leads to a larger drop in ability estimate than a slow incorrect response. One could argue, however, that when a subject guesses (fast response) he has less ability than a person with
a slow response who uses a strategy to calculate the answer but makes an error. The latter has at least some knowledge about how to reach the correct answer. In addition, Maris and van der Maas (2012) have proven that this scoring rule implies a 2 parameter IRT model and supported this model by data of the Amsterdam Chess Test. This enabled the derivation of an expected score from the scoring rule, which could be used in our Elo system.

However, more research concerning the HSHS scoring rule is needed. Above, we argue that an optimal trade-off between speed and accuracy exists. But it needs to be investigated whether subjects indeed understand this scoring rule correctly and whether they are able to optimally balance speed and accuracy. Other factors may also come into play. Based on the comments of teachers and children, we expect that the current implementation of the scoring rule into the games does not work optimally for some children. They might be distracted by seeing a coin disappear from the screen every second, making it difficult to focus their attention on solving the problem. Individual differences in impact of the scoring rule on performance need to be addressed in further research.

**Does Math Garden meet our aims?**

Much attention has been given in the Introduction (Chapter 1) to the aims of Math Garden. In the final part of this discussion we discuss whether our aims are met and which possibilities remain for further development. One of the first indications that Math Garden works in practice is the success of the tool in the educational market. Since the start of the spin-off company Oefenweb.nl in March 2009 the number of users of Math Garden has grown rapidly. September 2013 the Dutch version of Math Garden has more than 85,000 users, solving more than 500,000 math problems per day (see Figure 2). Math Garden has been granted three prices: 1) the Solberg-Verlinden/Wijnand Wijnen incentive prize for promising initiatives in the field of examination and testing (2009), 2) the IPON award for best innovative software product for primary education (2011), and 3) the Dutch e-learning Best Practice award (2011, awarded by the public). Below we will provide a more detailed discussion of our aims. We distinguish aims on student, teacher, and researcher level.
Student level

We assumed that practicing at one’s own level has a positive effect on children’s motivation and that a high success rate is an essential feature to achieve this effect. The results in Chapter 2 support our assumption that a CAT procedure with easy items has a positive affect on children’s motivation. Children played a lot outside school hours (33.2% of the answers) and children with low ability did not play appreciably less, indicating that motivation was similar for children of all abilities. In another study we further investigated the relation between Math Garden, motivation, math anxiety, perceived math competence, and math performance. The success rates of children were set at .6, .75, and .9. We found that the higher the success rate in Math Garden, the more children played the Math Garden games and the larger the improvement in math performance. We conclude, therefore, that the experience of success stimulates practice and that practicing math frequently at one’s own ability level improves math performance (Jansen, Louwerse et al., 2013). The latter conclusion is also supported by the study of Jansen, De Lange, and Van der Molen (2013) who found that adolescents (12-15 years) from special education who frequently played in Math Garden improved more in math ability than the control group who did not play in Math Garden. The sample size in the study was, however, small.

In the normal set-up of Math Garden children are free to choose between the three difficulty levels (success rate of .6, .75, or .9), allowing children to control the frequency of negative feedback. Hofman, Jansen, Visser, and Van der Maas (submitted) studied children’s choices and found that children in higher grades and with higher ability levels choose difficult items more often than children in lower grades and with lower ability levels. In addition, boys tend
to prefer more difficult items than girls. These results indicate that there are individual differences in preference for difficulty levels and that self-adaptive training and testing may be a valuable feature in individual training programs such as Math Garden. For further research, it might also be interesting to use the Math Garden data to study intra-individual differences in success rate preferences: do children switch frequently between success rates or do they prefer the same setting across games and times of the day/week?

**Teacher level**

On the teacher level we aimed to provide a measurement tool with diagnostic possibilities. Above we already discussed that Math Garden meets our aims of a valuable measurement tool for math abilities. Here we will discuss the more practical question whether Math Garden can provide valuable progress and diagnostic reports. In Chapter 2 we briefly discussed the diagnostic abilities of the program. In theory all kinds of reports about children’s results could be constructed as all answers of the children are logged. In practice we see two limitations. First, although Math Garden saves teachers time because all answers are automatically analyzed, they still have only limited time to study the results of their pupils. Second, we should keep in mind that teachers are not as trained in reading and interpreting result charts and tables as are researchers. Initially we developed reports for teachers that were much too detailed.

Currently we provide the progress reports in Math Garden that are displayed in Figure 3. First, a group overview (Figure 3a) is provided in which teachers can see how their pupils score on each math domain compared to children in the same reference group. This allows for a quick identification of pupils who score relatively low. In addition, it enables signaling which math games are relatively difficult for all pupils in the group. In this class overview teachers find links to individual results pages. On these pages teachers see individual developmental curves (Figure 3b) for each math game and an overview of items that are relatively difficult (called nightmares) and relatively easy (dream problems) for the child (Figure 3c), based on the difference between the expected score and the actual score of the child on recently answered items. The nightmares provide teachers insight into the needs for extra instruction of individual children.

Currently teachers still have to analyze the errors on the nightmare problems by themselves. Automatic analysis of children’s errors would be a big improvement. In Chapter 5 we developed a classification method for errors that could be implemented into Math Garden. Another request, often uttered by teachers, is insight into the achievement of learning goals. New learning goals for math and language skills (so-called reference levels; Dutch: referentieniveaus) were formulated for education in the Netherlands in the report Referen-
tiekader Taal en Rekenen (Meijerink et al., 2009). This report includes minimal and target levels for children of different ages. We are currently working on ways to provide these insights. The main challenge lies in the labeling of the problems that match the reference levels, as the report provides only general formulations of the minimal and target levels. In addition, as discussed above, Math Garden works with response times whereas the report is silent on the time allowed for solving problems. Another way of providing teachers more insight into the achievement of learning goals is reporting results for clusters of problems in each domain. Ideally, we would provide Elo ratings per sub group of problems. We aim to test this possibility in the future.

![Figure 3](image-url)

**Figure 3** Progress reports from Math Garden: school class overview (a), developmental curves per child per game (b), and nightmares and dream problems (c).
We are now also able to include instruction in Math Garden. We recently expanded the Math Garden system with instruction videos of the company Cédicu. By matching the labels of nightmare problems with labels of instruction videos we present each child with videos adjusted to the problems they find difficult. In the near future we will connect to other instruction material as well.

Finally, we believe that teachers themselves can help improve the Math Garden system. Now that the system is working, a main challenge when developing new math games is the construction of item banks. These should consist of at least hundreds of items of varying difficulty levels. Automatic item generation (Geerlings, Van der Linden, & Glas, 2013) is a possibility, using the results of Chapter 3 and 4. Information on which problem characteristics affect problem difficulty could be used to generate new items of desired difficulty levels. However, we also intend to build a system in which teachers can help improve the item banks by adding items that they believe are missing. The Elo system allows adding items into a working math game, as item difficulty estimates are updated after each time an item is answered, quickly leading to reliable difficulty ratings.

Figure 4 A schematic overview of the web application Mathsgarden.com (Rekentuin.nl), as discussed in Chapter 1. The parts of the figure with dashed lines represent possible extensions of the web application.
Figure 4 displays an overview of Math Garden and some of its possible extensions. Developing new learning tools for other school subjects or cognitive domains is one of these extensions. Indeed, the success of Math Garden has lead to the development of new products. In November 2012 the Language Sea (NL: Taalzee) has been launched, which already has more than 25,000 users. Language Sea is similar to Math Garden: it is an online learning environment for practicing and measuring language skills such as grammar, reading, and vocabulary. In the future, data from these learning tools could be combined for scientific research.

**Researcher level**

Last, but not least, we discuss the aims at the level of scientific research. In this thesis we have shown only a snapshot of the research possibilities with Math Garden and the Math Garden dataset. Since the start of this research project several researchers have become involved in the project and several grants have been awarded to research proposals concerning Math Garden and related systems.

One line of research concerns the effect of adaptive learning tools on children’s motivation and mathematical ability. The studies concerning children’s motivation have been discussed above. An important question that remains is whether training with Math Garden improves math ability. Unfortunately, this question is not easy to answer. In principle it requires a randomized controlled double blind experiment. Such experiments are very hard to implement in education. Kennisnet and Kohnstamm institute organized a small-scale experimental study and found some positive effects of Math Garden (Meijer & Karssen, 2013). The studies of Jansen, Louwerse et al. (2013) and Jansen, Van der Molen et al. (2013), which are described above, also found evidence for the positive effect of Math Garden on children’s mathematical abilities. These results are promising but replications in larger scale double blind experiments are required. Currently, we expect more benefits from randomized experiments within Math Garden. In these experiments key variables in the web application are manipulated in a controlled way, and assignment of users to conditions is random and, if possible, blind: without informing the users about the manipulation. In this way it is possible to test which features of the program are essential for stimulating playing frequency and for improvement of ability. The manipulation of success rates in the study of Jansen, Louwerse et al. (2013) is an example of such a randomized controlled experiment.

Another line of research concerns the analysis of item difficulties of the item banks of Math Garden, aiming to explain the item difficulties by a limited set of item characteristics. Chapter 3 and 4 are examples of such studies. Other studies have focused on the item diff-
ficulties of a task for logical reasoning, a deductive version of the well-known Mastermind game (Gierasimczuk, van der Maas, & Raijmakers, 2013); item difficulties of enumeration problems (Jansen et al., 2014); and item difficulties of a visuospatial working memory task (van der Ven, van der Maas, Straatemeier, & Jansen, 2013).

Finally, with Math Garden our aim was met to develop a measurement instrument that enables researchers to study mathematical development longitudinally, with high frequent measurements. Because the focus of this thesis shifted to the development of Math Garden, the study of the dynamics of mathematical development still requires further research.

Currently research projects at the University of Amsterdam focus on a number of questions regarding the dynamics of development, namely analyzing developmental change by modelling the time-series data of Math Garden (Young Talent grant, Van der Maas & Hofman, 2011) and regarding the mutual relations between cognitive and scholastic abilities. For example, Van der Ven et al. (2013) studied the relation between visuospatial working memory and mathematics for various math domains and across grades. Other researchers work on applications of the Math Garden technology to science learning (Brain & Cognition NWO project, Van der Maas & Van der Ven, 2011; Curious Mind Childrens Logic project, Raijmakers & van der Maas, 2011), psychometric questions about the scoring rule (Surf project, van der Maas & Klinkenberg, 2010), and psychometric innovations of the technology itself (Creative Industries project, Maris and van der Maas, 2012).

The extension of the Math Garden system to other learning domains enables researchers to study cognitive development with high frequent measurements in other domains. For the NWO science learning project of Van der Maas and Van der Ven a new learning environment is being developed, containing both games for executive functions (working memory, inhibition, planning) and higher mental abilities, such as reasoning and logic. Two of these games have already been pilotied in Math Garden, namely the balance scale task and the Mastermind game, which are discussed above. Potentially, we could also develop a game concerning children’s knowledge of the earth. In Chapter 6 we presented a cross sectional study concerning the developmental processes underlying children’s knowledge of the earth. Based on latent class analyses we concluded that this knowledge is fragmented and develops gradually as opposed to the assumption that development is stage-wise and that children develop different theory-like models at different ages. It would be interesting to study this developmental process longitudinally with high frequent measurements. The main challenge lies, however, in the development of an item bank for the earth game. In Chapter 6 we presented a new paper-and-pencil test with only 9 items, whereas hundreds of items are needed in a game using CAT. For other domains of the science learning environment this is easier to achieve. Gierasimczuk et al. (2013) developed an item bank of 321
items for the Mastermind game and the item difficulty of these items were well predicted by measures drawn from a logical task analysis.

A complete overview of papers resulting from the Math Garden project can be found in the Appendix.

**Bridging research and education**

Above we discussed the aims of Math Garden on different levels, arguing for its usefulness for both educational and scientific purposes. These two-fold aims make Math Garden a valuable tool, but also pose several challenges. The fact that Math Garden is used in everyday education solves the problem of finding participants for scientific research. Normally, finding schools willing to participate in a research project is very difficult as it puts an extra burden on already busy time schedules. With Math Garden schools participate in scientific research without requiring extra time of the teachers. The downside of the fact that Math Garden is used in everyday education is that assessment is much less controlled than in standard experiments. Children can use Math Garden anytime anywhere, so we cannot control for environmental influences. On the other side, Math Garden’s availability allows for the collection of a large dataset, which in turn makes it unlikely that environmental influences would have a large effect on the results. Combining results of Math Garden with that of controlled experiments is probably the best way to expand our knowledge on children’s development. An example of such combination of studies is the study of Jansen et al. (2014) concerning children’s skill to enumerate.

Summarized, Math Garden is the result of a unique combination of aims: the aim to investigate the dynamic development of cognitive abilities, the objective to improve psychometric measurement models and above all, the wish to motivate children to practice math to improve their abilities. The project shows that three diverse fields, cognitive development, psychometrics, and education, can be united and may even benefit from each other.