Self-adapting the success rate when practicing math

Jansen, B.R.J.; Hofman, A.D.; Savi, A.; Visser, I.; van der Maas, H.L.J.

DOI
10.1016/j.lindif.2016.08.027

Publication date
2016

Document Version
Final published version

Published in
Learning and Individual Differences

License
Article 25fa Dutch Copyright Act

Learning and Individual Differences 51 (2016) 1–10

Contents lists available at ScienceDirect

Learning and Individual Differences

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

Self-adapting the success rate when practicing math

Brenda R.J. Jansen [a,b,c,*], Abe D. Hofman [d], Alexander Savi [d], Ingmar Visser [a,b,c], Han L.J. van der Maas [b,c,d]

[a] Department of Developmental Psychology, University of Amsterdam, Mailbox 15916, 1001 NK Amsterdam, The Netherlands
[b] Cognitive Science Center Amsterdam, Nieuwe Achtergracht 129, 1018 WS Amsterdam, The Netherlands
[c] Yield, Research Institute of Child Development and Education, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands
[d] Department of Psychological Methods, University of Amsterdam, Mailbox 15906, 1001 NK Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 4 July 2015
Received in revised form 4 August 2016
Accepted 18 August 2016

Keywords:
Computer-adaptive testing
Self-adaptive testing
Mathematics
Math practice
Primary school

ABSTRACT

Use and benefits of the possibility to choose a success rate are studied in a math-practice application that is used by a considerable percentage of Dutch primary school children. Study 1 uses data that were collected with the application, using children's practice data ($N = 40{,}329$; grades 1–6). Children differed in their preference for a high, medium, or low success rate. Preferences were associated with gender, age, and ability, matching expectations that follow from the literature. Study 2 is an experimental study with 192 children from grades 3–6, using a pretest, training phase, and posttest. The possibility to choose a success rate was manipulated. Unexpectedly, beneficial effects for math practice, improvement of math skills, and self-belief concerning math were absent. Results suggest an appreciation of the possibility to choose, although beneficial effects of choosing were not observed for motivation to practice, skill improvement, and self-belief concerning math.

1. Introduction

Practice with math is, next to adequate instruction, essential for children to develop their math skills. Children themselves often bear the responsibility to practice regularly, but motivation to practice may fall short because practice can be boring when the difficulty level of problems is too low or frustrating when the level is too high.

In research on math anxiety, motivation to practice is considered to be one of the mechanisms explaining the negative relation between math anxiety and math performance (Ashcraft, Krause, & Hopko, 2007), which has been demonstrated repeatedly (e.g., Hembree, 1990; Ma, 1999). Math anxiety concerns behavioral, affective, physical, and cognitive complaints when solving math problems or anticipating doing math (Ashcraft & Kirk, 2001). The behavioral component is assumed to involve the avoidance of math (Ashcraft & Krause, 2007; Hembree, 1990). A negative spiral may occur when a child develops a negative attitude regarding math after failing at difficult math problems and refrains from doing math in the future, missing essential math practice (Ashcraft et al., 2007). Although the majority of children will not develop an aversion as strong as math anxiety, failing at math problems may still demotivate the practice of math.

The present investigation is motivated by the question whether principles of computer-adaptive and self-adaptive testing, in which difficulty of problems is adapted to the child's ability, can be helpful in promoting the motivation to practice math. More specifically, when difficulty of problems is adapted to the child's ability, success in math can be manipulated and adapted to both the cognitive possibilities and emotional needs of the individual child.

We examine this question by conducting two studies with Math Garden. Math Garden is a web-based application for monitoring and practicing math skills, based on computer-adaptive testing and developed by the University of Amsterdam (Klinkenberg, Straatemeier, & van der Maas, 2011). A considerable percentage of schools in the Netherlands (5%; over 120,000 users) employ Math Garden in their daily school programs and allow pupils to practice their math skills with the application. The resulting data about practice, skill, and particular choices by the children are central in Study 1. These data have high ecological relevance because they reflect the preferences and skills of a significant part of the Dutch school population, while they are practicing in their natural school setting. It follows that there is only limited control over the conditions under which participants use the system and that the use of Math Garden in schools makes it difficult to perform experimental research. However, an experimental set-up is possible in an isolated sample (of limited size), such as in Study 2. In Study 2, the question whether the possibility to choose a success rate positively affects practice and skills is investigated. Together, Studies 1 and 2 provide insight into the possible beneficial effects of principles of computer-adaptive and self-adaptive testing on math practice and development of math skills.

Adapting the difficulty level of problems to the individual's ability may benefit children's motivation to practice. Adaptivity is a key feature of computer adaptive testing (CAT; Eggen & Verschoor, 2006; Van der

* Corresponding author at: Department of Developmental Psychology, University of Amsterdam, Mailbox 15916, 1001 NK Amsterdam, The Netherlands.
E-mail address: b.r.j.jansen@uva.nl (B.R.J. Jansen).

http://dx.doi.org/10.1016/j.lindif.2016.08.027
1041-6080/

Linden & Glas, 2000; Weiss, 1982; Wainer et al., 2000). Correctly answered problems are followed by more difficult ones and incorrectly answered problems are followed by easier ones. Usually, problem difficulties are determined beforehand in a separate sample. Aims of testing of course differ from those of practice. Testing should be efficient and administration of each problem should yield as much information about a student's ability as possible. In most CATs this is accomplished by selecting problems with difficulties that match a student's current estimated ability (Birnbaum, 1968; Eggen & Verschoor, 2006), that is problems with a probability of success of 0.5. This way the length of the test is minimized. For practice purposes, length of the test should be weighed against motivation and pleasure of students. Obtaining only 50% correct may be a challenge for few but a frustration for many. Although based on CAT principles, Math Garden uses much higher success rates, of on average 75%. Jansen et al. (2013) applied an experimental design in Math Garden by assigning children (8–13 years) to conditions in which they practiced at a fixed success rate. They observed that practice increased with increasing success rate, and that improvement of skills again increased with increasing practice. Hence, high success rates seem beneficial for math practice.

Math Garden also uses principles of Self-Adaptive Testing (SAT; Rocklin & O'Donnell, 1987), which allow participants to select a success rate. SAT was designed because it was doubted that success rates of 50% were "motivationally optimal for all examinees" (Rocklin & O'Donnell, 1987, p315; see also Ortner & Caspers, 2010). In SAT, an individual selects a success rate and SAT thus accommodates possible individual differences in preferred success rates. A problem of SAT is that predetermined success rates are used for all individuals, regardless of an individual's ability. However, what's difficult for some may be easy for others.

Restricted Self-Adaptive Testing (RSAT; Wise, Kingsbury, & Houser, 1993) solves this issue because an initial estimate of the individual's ability is used for the composition of individual problem sets for each success rate. Difficulty of problems may differ across individuals but do result in similar success rates for a selected success rate. Hence, RSAT not only allows a student to select a success rate, but also secures that problems presented result in a comparable success rate for other individuals who may differ in ability but select the same success rate. Roos, Wise, and Finney (1998) compared RSAT to SAT and CAT and found that measurement precision was comparable to that in CAT. Math Garden allows students to select problem difficulty levels in a way that is comparable to that in RSAT although technical implementation differs importantly (Klinkenberg et al., 2011). Success rates in Math Garden are low, medium, and high (aimed success rates of 60%, 75%, and 90%, respectively).

The first aim of Study 1 is to explore children's use of the possibility to select a preferred success rate in Math Garden. It is expected that children will differ in their selections because individual differences in preferred success rates are likely in learning settings (e.g., Meyer, Turner, & Spencer, 1997). Moreover, selections of success rates may be importantly related to individual differences in the tolerance for the negative feedback that often comes with making mistakes. The second aim of Study 1 is therefore to relate preferences for success rates to a set of background variables that have demonstrated relations to individual differences in the tolerance for negative feedback: gender, ability, and age. Concerning gender, it is observed that females are more sensitive to punishment (Cross, Copping, & Campbell, 2011; Moeller & Robinson, 2010), pay more attention to negative stimuli (Li, Yuan, & Lin, 2008; Schirmer, Zysset, Kotz, & von Cramon, 2004; Yuan et al., 2009), and tend to avoid negative feedback to a higher extent (Weller, Levin, & Bechara, 2009), compared to males. It has been suggested that females are more concerned with pleasing adults, which might make them more anxious for encountering failures (Pomerantz, Altermatt, & Saxon, 2002). Individual differences in tolerance for negative feedback have been associated with ability as well. Hontangas, Ponsoda, Olea, and Wise (2000) detected a failure-tolerant strategy: Students using this strategy selected a problem of the same difficulty level after an error. Remarkably, these students had higher ability levels than students using other strategies. The finding suggests that good students prefer a challenge and therefore select more difficult items that is, lower success rates. The other way around, students who are able to cope with failures may be more likely to perform at a higher level. However, Revuelta (2004), who detected the failure-tolerant strategy as well, found no clear relation between this strategy and ability. The relation between ability and preferred success rate can be studied well with Math Garden data because they result from RSAT: If high ability students select a high success rate (relatively easy items), they are likely to be presented with items of overall average difficulty. Also, lower ability students selecting a low success rate (relatively hard items) might be presented with the same problems of overall average difficulty. Hence, the relation between ability and preferred success rate can be studied, independent of the difficulty of the items. Individual differences in tolerance for negative feedback have been related to age as well. Learning studies show that children up to about twelve years perform more inaccurately after negative feedback than after positive feedback (Van Duijvenvoorde, Zanolie, Rombouts, Raijmakers, & Crone, 2008). Perhaps, children's preference for positive feedback over negative feedback decreases with age, accompanied by a decreasing preference for high success rates. In sum, individual differences in tolerance for negative feedback are numerous in the literature and are to be expected in the preference for success rates in Math Garden.

Study 2 is focused on possible beneficial effects of self-selecting practice problems based on likely success rates on the motivation to practice and the development of skills. It has been demonstrated that tests using SAT principles reduced the influence of test anxiety on test performance, compared to CAT (Hontangas, Olea, Ponsoda, Revuelta, & Wise, 2004; Pitkin & Vispoel, 2001; Rocklin & O'Donnell, 1987; Vispoel & Coffman, 1994) although not consistently (Johnson, Roos, Wise, & Plake, 1991; Ponsoda, Wise, Olea, & Revuelta, 1997; Shermis, Mzumara, & Bublitz, 2001; Vispoel, Rocklin, & Wang, 1994). Also, it is shown that ability estimates are higher in SAT than in CAT (e.g., Vispoel et al., 1994). Studies showing beneficial effects of RSAT are less numerous (but see Roos et al., 1998). The beneficial effects of SAT on test anxiety and ability may originate from a higher focus on the task, which may be effectuated through the selection process in SAT. RSAT may also be beneficial because it is in line with the need for autonomy, one of the central concepts of self-determination theory (Ryan & Deci, 2000; for an overview of their work: Niemiec & Ryan, 2009). Selecting a difficulty level in RSAT implies choice, which increases autonomy (Katz & Assor, 2007). Changing a success rate directly influences the required effort in the task (Reeve, Nix, & Hamm, 2003), and is therefore highly relevant to students' interests and goals. These findings all suggest that implementing RSAT in a math practice application, with choice as a result, may be beneficial for children's motivation to practice and for their math skills. The aim of Study 2 is to test, in an in vivo experimental design, whether the option to select a success rate increases the motivation to practice math and improves math performance.

Summarized, in Study 1, the use of self-selecting the success rate is explored by analyzing a large data set that was collected with Math Garden. Regular conditions of Math Garden are used, in which all children have the opportunity to select a success rate. Next, it is investigated whether individual differences in preferred success rate relate to gender, age, and ability. We expect that boys tend to select a lower level of success rate than girls and that preferred success rate decreases with ability and age. Results of Study 1 are a precondition for Study 2; an experimental design is used to investigate the impact of self-selecting the success rate on practicing math and math performance. The default condition that entails choice of success rate is contrasted with conditions with fixed success rates with easy, medium, or hard problems. We expect that children in the choice condition practice more and improve their abilities more than children in the fixed conditions. This design also offers the opportunity to replicate the finding of Jansen et al. (2013), i.e., that practice increases with increasing success rate. For this aim, we compare the three fixed conditions and expect that success rate is positively associated with math practice, which is again positively associated with improvement in math performance.

Also in Study 2, we explore whether success rate increases self-beliefs concerning math. Self-belief relates to self-concept (assessment of one's ability in math; e.g., Marsh & Martin, 2011) and self-efficacy (the degree that one is confident of his or her ability to accomplish a math task well; Usher & Pajares, 2008, 2009). Self-beliefs concerning math are reciprocally related to math performance (e.g., Erturan & Jansen, 2015; Marsh & Martin, 2011; Pajares, 1996; Usher & Pajares, 2006; Williams & Williams, 2010). Experiencing success in math may positively affect self-beliefs concerning math.

2. Study 1

2.1. Material and methods

2.1.1. Participants

In principle, participants were all children in grades 1–6 who had a subscription for Math Garden. Schools and families can subscribe to Math Garden. Upon subscribing, schools and families were asked for permission to use their data for scientific publications. Schools take on the responsibility to inform the parents of their pupils about use of data and voluntary participation. Data from children whose parents did not give permission were excluded from the dataset.

For inclusion in the current dataset, it was required that children had finished >45 problems in the addition game of Math Garden between September 2012 and September 2013. The addition game was selected because the game was presented to children of all grades. The current Math Garden data set contained data of 40,329 participants (Grade 1: $N = 3031$ participants, M age $= 6.70$ years, $SD = 0.72$; Grade 2: $N = 7360$, M age $= 7.46$ years, $SD = 0.63$; Grade 3: $N = 7729$, M age $= 8.51$, $SD = 0.61$; Grade 4: $N = 7447$, M age $= 9.47$ years, $SD = 0.61$; Grade 5: $N = 6964$, M age $= 10.49$ years, $SD = 0.62$; Grade 6: $N = 7798$, M age $= 11.61$, $SD = 0.64$). Percentage of females ranged from 45% to 50% in the different grades. The first 15 items were omitted since children still had to get acquainted with the task, resulting in a minimum number of 30 completed problems available for analysis. The average number of solved problems in the addition game was 202 ($SD = 229$, range $= 30$–9500).

2.1.2. Math Garden

2.1.2.1. Interface. The central page in Math Garden was the individual's personal garden (left panel of Fig. 1), containing flower beds connected to math games. At the central page, a participant could select a success rate and select a game to practice. When entering Math Garden for the first time, the default success rate was Medium, represented by a smiley with two drops of sweat. Children could change the success rate to High by clicking the smiley with one sweat drop, which made the problems easier, or to Low by clicking the smiley with three sweat drops, which made the problems more difficult (see Fig. 1). The selected success rate applied to all problems in a game. The preset aimed success rates of the high, medium, and low success rate were 90%, 75%, and 60%, respectively. Children returned to the central page after finishing a session of the game or premature abortion of a session. At the central page, the smiley of the last selected success rate was still highlighted. The success rate could be re-adjusted before starting a new session of a game.

In the addition game, a problem consisted of a bare addition problem (see right panel of Fig. 1 for an example). Problems varied from easy (e.g., $1 + 1$) to hard (e.g., $36.40 + 86.09$). Children answered by choosing one from six response options presented on the right of the problem. A session of a game consisted of fifteen problems presented sequentially. Participants had 20 s to solve each problem, represented by 20 coins at the bottom of the screen. A coin disappeared with each expiring second. After giving the correct answer, participants were rewarded a virtual coin for every second that was left. Participants lost this number of coins from their collection if they responded incorrectly. Coins were worth ½ point in the level with high success rate, 1 point in the level with medium success rate, and 2 points in the level with low success rate. Varying the value of the coins is done to create similar rewards in coins between the success rate levels. Premature abortion of a game was possible but not rewarded, whereas a fully completed session was rewarded with 25 extra coins. Coins were used to buy virtual trophies (e.g., medals, pennants), to be placed in a personal trophy cabinet.

2.1.2.2. Implementation of restricted self-adaptive testing in Math Garden. In Math Garden, restricted adaptive selection of math problems was implemented by scoring both problem difficulties and children's abilities on the same scale. The higher the score, the more difficult a problem or the higher the child's ability. When the child's ability was very low and the problem's difficulty was very high, the problem would be too difficult for the child. When the child's ability and the problem difficulty were close together, probability of solving the problem correctly would be around 50%. The higher the difference between child's ability and problem difficulty, the higher the probability of solving the problem correctly, in limited time. The exact probability was estimated using Item Response Theory (IRT).

At the very start of Math Garden, problem difficulties were unknown and were given initial values, based on problem size. Children's abilities were unknown as well and initial values were based on age only. Both difficulties and abilities were adjusted after children started solving problems. Problem difficulty increased when a problem was solved incorrectly and/or for which solving time was long, and decreased when a problem was solved correctly and even more when it was solved fast. Also, children's abilities were adjusted based on their success and solving times. Adjustment was based on the Elo (1978) rating system, designed for chess competitions. Klinkenberg et al. (2011) and Maris and Van der Maas (2012) provide technical details. Adjustment of difficulties and abilities was an ongoing process. Although estimates of

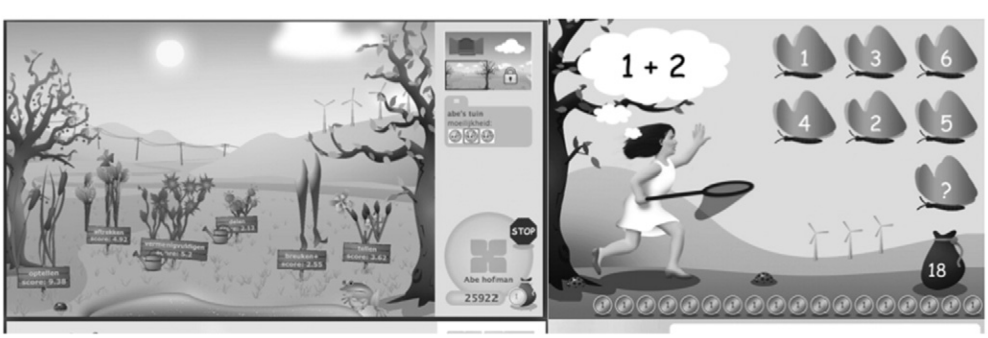

Fig. 1. Example of Math Garden. Left screen shows the central page. Right panel shows an example of an addition problem.

problem difficulties had converged to stable levels, estimates of children's abilities were constantly updated.

Selection of a subsequent problem for an individual child was based on the difference between the child's estimated ability and the problem difficulties and on the selected success rate. If the child selected the high success rate, the difference between ability and difficulty was required to be larger than when the child selected the medium success rate, and this difference was again required to be larger than when the child selected the low success rate.

2.1.2.3. Data. Data in Study 1 comprised the selected success rate for each attempted problem, except for the first 15 items since children had to get acquainted with the task. Schools or families provided background information on grade and gender upon subscribing. Estimated ability in Math Garden's addition game was used as an indicator of addition performance. Finally, switches between success rates were inspected to check whether children were aware of the possibility to select a success rate.

2.1.3. Procedure

Participants received a login and password with which they logged onto the website of Math Garden. Participants could play at school when the school permitted so or required to and participants could play at home. Upon logging in, participants arrived at their personal garden. Math Garden was self-explanatory. Online help was available, but the majority of the students learned to understand the procedure of choosing a game, selecting a success rate and the principles around feedback by themselves. Success rates were not communicated explicitly, but the value of coins, halved for the high success rate and doubled for the low success rate, were clearly visible.

2.2. Results

2.2.1. Manipulation checks

Preset success rates are only reached if the adaptive algorithm operates optimally. Actual success rates were 79%, 72%, and 62% for levels high, medium, and low, respectively. Success rates of the high and medium success rates were lower than the aimed rates of 90% and 75%, and the success rate of the low level was higher than the aimed rate of 60%. However, the success rates clearly differed from each other, χ^2 (2, $N = 8.2 \times 10^6$) $= 2.1 \times 10^5$, $p < 0.001$.

2.2.2. Categorization of preferences for success rates

Since number of attempted problems differed between children, proportions of selecting each success rate were computed per child. Table 1 shows the average proportions of the selections of the high, medium, and low success rate, by grade. A general pattern of decreased preference for the high success rate and increased preference for the low success rate with increasing age emerged.

The proportions of playing at the high, medium and low level of success rate were subjected to iterative cluster analysis. In iterative cluster analysis, multivariate data are used to categorize participants into a limited number of clusters. Individuals with similar preferences for success rates were clustered and separated from individuals in other clusters with dissimilar preferences for success rates. The goal of this analysis was to classify each subject to a cluster representing the overall preference for the success rates. The exploratory data reduction approach of cluster analysis suits the purpose since it allows for finding clusters representing each success rate, but also clusters representing combinations of success rates (e.g., 50% high and 50% medium). It reveals more information than the average success rate because the same average (e.g., 75%) can be achieved with different strategies (e.g., switching between low and high success rate, but also consistently playing on medium). Cluster analysis may also be more informative than the dominant success rate when a child plays as much at one level as at the other.

Solutions with 2 to 10 clusters were estimated and compared by inspecting the total within cluster sum of squares, which was set out against the number of clusters in a line graph. In general, the total within cluster sum of squares lowers with increasing number of clusters. A sharp bend in the line graph indicates the appropriate number of clusters. Three clusters turned out to be a good balance between parsimony and precision of description. Table 2 shows, for each cluster, the average proportion of attempted problems at each success rate, the number of participants, and the average number of switches. Sizes of the clusters were comparable. The choices in each cluster concentrated mainly around one success rate and the clusters were therefore referred to as “High success rate”, “Medium success rate”, and “Low success rate”. Switching was present in all clusters. The three-cluster solution evidenced individual differences in preference for a success rate. Switches showed that children used the possibility to customize the success rate in the current low-stakes math practice situation.

2.2.3. Association of preferred success rates with age, gender, and ability

Ability in math refers to performance in comparison to peers because math skills in school are related to age (Krinzinger, Kaufmann, & Willmes, 2009; Van der Ven, Kroesbergen, Boom, & Leseman, 2012). Centering by age group shows relative performance and is a method to control for this age effect. Here, we used z-scores for centering, resulting in an average standard score of 0 in all grades.

A multinomial regression analysis was performed with number of switches, gender, centered ability estimates, and grade as independent variables and cluster as the dependent variable. Only a main effect was estimated for number of switches but interactions were allowed between gender, ability, and grade. We used the so-called backwards method for model selection, using BIC as a fit measure. In multinomial logistic regression, standardized b signals the change in probability of being in a given category compared to being in a reference category, associated with a change of one standard deviation in the independent variable. Here, the cluster with Medium as the preferred success rate is the reference category.

The selected model contained a main effect of switches (High success rate compared to Medium: $b = 0.10$, $p < 0.001$, and Low success rate compared to Medium: $b = 0.07$, $p < 0.001$), evidencing that a higher number of switches was associated with a higher probability of being in clusters with a preference for a high or low success rate, as compared to the cluster with a preference for the medium success rate. This effect matched the fact that the medium level is the default level in Math Garden. The selected model also contained significant main effects of gender, ability, and grade but all were qualified by two-way interactions

Table 1
Proportion of attempted problems at each success rate by grade in Study 1.

Proportion of problems (standard deviations between parentheses)			
Grade	High success rate	Medium success rate	Low success rate
1	0.45 (0.50)	0.37 (0.48)	0.18 (0.38)
2	0.49 (0.50)	0.31 (0.46)	0.20 (0.40)
3	0.45 (0.50)	0.30 (0.46)	0.25 (0.43)
4	0.37 (0.48)	0.31 (0.46)	0.32 (0.47)
5	0.28 (0.45)	0.31 (0.47)	0.41 (0.49)
6	0.20 (0.40)	0.32 (0.47)	0.48 (0.50)

Table 2
Three-cluster solution of selections of success rates in Study 1.

	Cluster		
	High success rate	Medium success rate	Low success rate
p(high)	0.71	0.09	0.10
p(medium)	0.16	0.81	0.14
p(low)	0.13	0.10	0.76
N (%)	12,865 (32%)	14,603 (36%)	12,861 (32%)
Mean #Switches	5.26	2.99	4.29

between ability and gender (High success rate compared to Medium: $b = 0.01$, *ns*, and Low success rate compared to Medium: $b = -0.05$, $p < 0.001$), ability and grade (High success rate compared to Medium: $b = -0.05$, $p < 0.001$, and Low success rate compared to Medium: $b = 0.01$, $p = 0.001$), and grade and gender (High success rate compared to Medium: $b = 0.03$, *ns*, and Low success rate compared to Medium: $b = -0.09$, $p < 0.001$). The interaction effects are illustrated in Fig. 2. For ease of display, only probabilities for clusters with a preference for the high or low success rate, for grades 1, 3, and 5 were displayed in Fig. 2. Probabilities of the clusters sum to one, and the probability of being in the cluster with a preference for the medium success rate is the complement of the summed probabilities of the clusters with a preference for a high and low success rate. Grade is a continuous variable in this analysis so the results of grades 2, 4, and 6 follow the pattern of the nearest grades displayed.

The interaction between ability and gender is shown in the upper left panel of Fig. 2, which shows that the positive relation between ability and the probability of being in cluster "Low success rate" was stronger for males than for females. The interaction between ability and grade is demonstrated in the upper right panel of Fig. 2 and indicated that the main effect of ability was stronger in higher grades. That is, the probability of being in the cluster with a preference for a high success rate decreased and the probability of being in the cluster with a preference for a low success rate increased with increasing ability, but these relations were stronger in higher than in lower grades. The interaction between grade and gender is demonstrated in the lower panel of Fig. 2, which shows that the difference between boys and girls, in the probability of being in the cluster with a preference for a low success rate, was larger in higher grades.

Summarized, the results of Study 1 matched the expectations: children differed in their preference for a likely success rate and these individual differences related to gender, grade, and ability. Children with higher ability showed a preference for the low success rate. This negative relation was stronger for boys than for girls and strongest in the highest grades. The difference between boys' and girls' preference for the low success rate, with boys preferring the low success rate more than girls, was larger in the higher grades. The gender difference matches the findings in studies on feedback learning that girls are more sensitive than boys to negative feedback (e.g., Cross et al., 2011; Moeller & Robinson, 2010). The present finding that high ability is related to higher tolerance for negative feedback matches the findings by Hontangas et al. (2000) but not those of Revuelta (2004). Finally, the negative association between age and success rate is consistent with findings from feedback learning studies (Van Duijvenvoorde et al., 2008).

3. Study 2

Study 1 led to the conclusion that children from primary school used the option to choose a likely success rate when practicing math with Math Garden. Also, individual differences in preferred success rate matched hypotheses on the relation with gender, grade, and ability, which were derived from the literature on tolerance for negative feedback. Study 2 is set up to investigate 1) whether choice of success rate is beneficial for math practice and math performance; 2) whether success rate is positively related to math practice and math performance (an attempt to replicate findings of Jansen et al., 2013); 3) whether success rate is positively associated with improvements of self-belief. Study 2 uses a pretest-posttest design with an experimental period in between in which children practiced math in Math Garden. Different conditions were applied in which children either had the option to choose a success rate themselves or practiced at a fixed success rate. Math skills and self-belief were assessed at pretest and posttest. Number of attempted problems and time spent in Math Garden were logged and used as indicators of math practice. It was expected that 1) children in the choice condition practiced more and improved more in math performance; 2) math practice increased with increasing success rate and that math practice and improvement in math performance were positively related; 3) improvement in self-belief measures increased with increasing success rate.

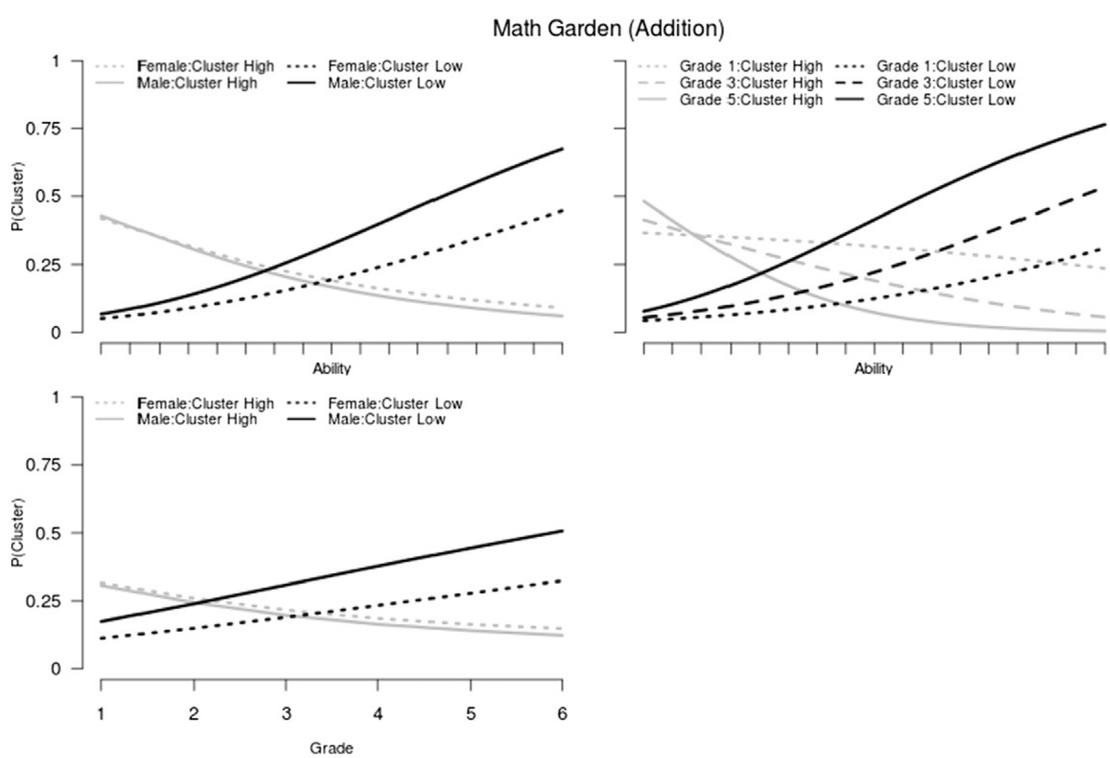

Fig. 2. The upper panels show the relation between ability and probabilities of clusters with a preference for a high and low success rate, by gender (left panel) and by grade (right panel). The lower panel shows the relation between grade and probabilities of clusters with a preference for a high and low probability, by gender. The cluster with a preference for the medium success rate is the reference category and is left out for clarity.

3.1. Materials and method

3.1.1. Participants

The initial sample consisted of 192 students from grades 3–6 of a primary school in the Netherlands. Parents received an information letter and could indicate that their child was not allowed to participate in the study. The departmental ethics committee approved of the procedures. There were 50 children from grade 3 (46% girls, *M* age = 8.42 years, *SD* = 0.35), 50 children from grade 4 (36% girls, *M* age = 9.50, *SD* = 0.42), 54 children from grade 5 (56% girls, *M* age = 10.56, *SD* = 0.45), and 38 children from grade 6 (47% girls, *M* age = 11.66 years, *SD* = 0.55.

Children were assigned randomly, across classrooms, to four conditions (Choice, High success rate, Medium success rate, Low success rate). There were 48 children in each condition but data from one female fourth-grader, who was assigned to the condition with low success rate, were excluded from analyses because she did not practice at all in Math Garden. Before implementing the intervention, conditions were compared on gender and grade distribution, math skills and self-belief scores. Conditions did not differ with regard to grade distribution, χ^2 (9) = 0.31, $p = 1.00$, or gender distribution, χ^2 (3) = 4.03, $p = 0.259$. Also, conditions did not differ with regard to math skills scores and self-belief scores, all $p \geq 0.275$.

Three children (one from grade 4, one from grade 5, and one from grade 6) were absent during the pretest, whereas 7 children (2 from grade 3, 4 from grade 5, and 1 from grade 6) were absent during the posttest. Three children did not fill in the math skills test at either pretest or posttest due to waning motivation or planned leave (2 in grade 3, 1 in grade 4). Four children skipped one or more pages of questionnaires at either pretest or posttest (3 in grade 3, 1 in grade 5). If a child missed one or more items of a measure, data of this child were omitted from the analysis concerned.

3.1.2. Design

The study consisted successively of a pretest, an experimental period, and a posttest. A math test and self-belief measures were administered at both pretest and posttest. After the pretest, children were assigned to either the Choice condition, in which success rate depended on the success rate children selected themselves, or a fixed condition, in which success rate was aimed to equal 90%, 75%, and 60%, respectively. On average, children's success rate in the choice condition was 76% (*SD* = 0.07). Success rates in the fixed conditions were on average 84% (*SD* = 0.04), 77% (*SD* = 0.05), and 71% (*SD* = 0.05), for the conditions with high, medium, and low success rate, respectively. Success rates of the fixed conditions differed significantly from each other, all p's < 0.001.

3.1.3. Material

3.1.3.1. Math Garden. All children always had access to the basic garden, which had plants linked to the games for addition, subtraction, multiplication, division, fractions, and counting and a game concerning correct sequence of operations. In the addition and subtraction games, a problem was presented with six response options. All other games used a "fill-in-the-blank" with a number pad. A bonus garden opened after a child had played all games in the basic garden. This bonus garden is part of the Math Garden application. It was designed to increase children's interest in practicing math. The garden includes games that are related to arithmetic, but children may find these more fun than the regular math games. Examples of games in the bonus garden are reading the time and a memory game. Children only have access to the bonus garden if they have practiced sufficiently in the basic garden. Hence, children who practice more in the basic garden have earlier access to the bonus garden. Settings and content of the bonus garden were equal for all conditions.

The regular version of Math Garden, described in Study 1, matched condition Choice in Study 2. See Material and methods section of Study 1 for details on this condition. The Choice condition, in which children could select the success rate themselves, was contrasted with three fixed conditions in which success rate was fixed to high, medium, or low. As in Study 1, coins were worth 1/2, 1, or 2 points in the conditions with high, medium, and low success rate, respectively.

The application logged each attempted math problem, the success rate, date and time of problem presentation, child's response, response time, and accuracy, for each child. From these data, we constructed the variable Time on task, which was equal to the number of problems x average response time. Number of attempted problems and Time on task were partitioned between the high, medium, and low success rate for children who were allowed to choose a success rate.

3.1.3.2. Math skills. The Tempo Test Automatization (TTA; De Vos, 2010) is used for an estimation of the degree to which addition, subtraction, multiplication and division facts have been automated. TTA is suited for children in grades 1–6 of primary education. The test consists of four subtests, each consisting of 50 math problems. Participants have exactly 2 min to complete as many math problems as they can, per subtest. The score is the number of correct answers, with a maximum of 50 for each domain and a maximum of 200 for the full test. A ceiling effect occurred in the domains of multiplication and division in grades 7 and 8 on the pretest. Hence, we only used the scores for the domains addition and subtraction. Addition and subtraction scores were converted into z-scores by grade (see Study 1 for justification of using z-scores). To study representativeness of the sample, the grade averages of the total of addition and subtraction scores were converted into percentiles (De Vos). Average scores fell into the 58th, 56th, 50th, and 50th percentile for grades 5 to 8, respectively, showing that the present sample was representative.

3.1.3.3. Self-belief. The sub-scale "Self-efficacy" of the Global Motivation Questionnaire (Prast & Van de Weijer-Bergsma, 2013) consists of seven questions. Example questions are "Can you often solve all math problems correctly?" and "Do you think you will understand all math problems that you will be taught?". Children answered the questions on a four-point scale ranging from not true (1 point) to completely true (4 points), making the total score range from 7 to 28. Internal consistency of this scale was $\alpha = 0.83$ for the pretest in the present study.

The Sources of Middle School Mathematics Self -Efficacy Scale (SMSMSS; Usher & Pajares, 2009) consists of 24 statements and provides insight into the degree of self-efficacy. It is suited for children between seven and twelve years. Statements can be divided into four scales. Example statements are: "I make excellent grades on math tests" (mastery experience), "Seeing kids do better than me at math, pushes me to do better" (vicarious experience), "I have been praised for my ability in math" (social persuasion) and "Doing math work takes all of my energy" (physiological state). Children indicated how much the statement was true for them on a 6-point scale, ranging from 1 (not true at all) to 6 (completely true), with the total score ranging from 24 to 144 points. The scale was translated into Dutch for the present study. A principal component analysis resulted in the extraction of four factors with an eigenvalue higher than 1, which were Varimax-rotated. Together, the factors explained 58% of the variance. Statements that referred to mastery experience all loaded on the same factor, just like statements referring to social persuasion and statements referring to physiological state. However, one statement ("I can imagine that I can solve difficult math problems correctly"), which should refer to indirect experience, also loaded high on the factor associated with mastery experience. Finally, all statements referring to indirect experience, except the statement above, loaded on the same factor. Hence, the factor structure highly resembled the solution of Usher and Pajares and the factors were associated with mastery experience, vicarious experience, social persuasion, and physiological indexes. Internal consistency for the scores collected at the pretest was $\alpha = 0.91$.

The sub-scale "*Academic self-concept math" of the Self Description Questionnaire* (SDQ I; Marsh, 1988; translated to Dutch by Simons & Fisette, 2001) consists of six statements. Example statements are "I look forward to math lessons", "I learn arithmetic fast". Children indicated the degree to which the statement applied to them on a six-point scale, ranging from not true at all (1 point) to completely true (6 points), with the total score ranging from 6 to 36 points. Internal consistency of the scale was $\alpha = 0.88$ for the pretest in the present study.

The Perceived math competence scale was an addition to the Dutch translation of the Perceived competence scale by Harter (1982); Dutch translation by Veerman, Straathof, Treffers, Van den Bergh, & Ten Brink, 2000; introduction of Perceived math competence scale by Jansen et al., 2013) and consists of six pairs of descriptions. Each description referred to a group of children. Children indicated the group they felt they belonged to most and next indicated whether the description was very true (1 or 4 points, depending on the selected group) or a little true (2 or 3 points, depending on the selected group). An example of a pair of descriptions is "Some kids often forget math strategies/Other children can easily memorize math strategies". Scores ranged from 6 to 24 points on the Perceived math competence scale. Internal consistency for the scale was $\alpha = 0.74$ for the pretest in the present study.

3.1.4. Procedure

The pretest and posttest consisted of a group-wise administration of the subscale Self-efficacy, SMSMSS, subscale Academic Self-concept math, Perceived math competence scale, and the TTA (math skills), in this order. Administration of the questionnaires, including instructions, took approximately 20–30 min. Administration of the TTA, including instructions, lasted about 10 min.

The experimental period started a week after the pretest. Children received a login name and a password, providing them access to Math Garden. Teachers received this information too, as well as a brief written explanation of Math Garden. Although Math Garden is self-explanatory, the experimenter, a trained master student, explained Math Garden to the children in their classrooms. She told children that they were in different groups and that value of the coins differed between groups.

The experimental period lasted on average 7.5 weeks ($SD = 0.28$). The period of 7.5 weeks is comparable to the duration of activity plans in the Netherlands, in case learning lags behind and after which it is decided whether the activities were effective for the child. In the first week of the experimental period, parents received a letter through their children in which they were asked to encourage their children's practice with Math Garden. The experimenter monitored frequency of practice through the system. She sent a number of emails to the teachers, offering assistance with Math Garden and informing about the degree of practice. In addition, she repeatedly visited the school to motivate the students.

3.2. Results

3.2.1. Correlations between self-belief questionnaires, and with math performance

Correlations between scores on the self-belief questionnaires and standardized scores on the math test were conducted to obtain an indication of the validity of the self-belief questionnaires. Scores of the self-belief questionnaires were standardized by grade because scores of children in grade 3 tended to be higher than those of children in grade 5 (and sometimes 6) for all self-belief measures, except SMSMSS. Correlations between the self-belief questionnaires and the math test ranged from weak to moderate, the lowest being $r = 0.32$, $p < 0.001$ (posttest: TTA and Academic self-concept math scale) and the highest being $r = 0.48$, $p < 0.001$ (posttest: TTA and subscale Self-efficacy). Correlations between the self-belief questionnaires ranged from $r = 0.51$, $p < 0.001$ (pretest: Academic self-concept math and Perceived math self-competence) to $r = 0.77$, $p < 0.001$ (posttest: Self-efficacy and SMSMSS), supporting the assumption that the various instruments had a common component, probably self-belief.

3.2.2. Conditions compared on motivation to practice

Table 3 shows the Number of attempted problems and Time on Task for the experimental period, split by basic and bonus Math Garden (see Method), by condition. The distributions of both variables were positively skewed. The minimum of attempted problems was 37 in the basic garden. Table 3 suggests that children in the condition with a low success rate attempted less problems than children in the condition with a medium success rate. However, the average RT for solving a problem in the condition with low success rate was 8664 ms (averaged over basic and bonus games), whereas solving a problem in the condition with a medium success rate took 7767 ms (condition High: 6301 ms; condition Choice: 7779 ms). Possibly, children in the condition with a low success rate spent a comparable time in the Math Garden but were only able to attempt a smaller amount of problems in that time, compared to other conditions.

A Multivariate Analysis of Variance (MANOVA), with factor condition as independent variable (4 levels: Choice, High, Medium, Low) and time on task and number of attempted problems as dependent variables showed a multivariate effect of condition, $F(6, 374) = 14.34$, $p < 0.001$. Table 3 shows that both number of attempted problems and time on task were highest in the condition with medium success rate. The choice condition also showed a high number of attempted problems and time on task. Univariate tests however did not show any significant differences in number of attempted problems and time on task between conditions. Non-parametric analyses were performed as well because assumptions for MANOVA did not always hold. Although the p-value was at trend level for number of attempted problems, the non-parametric analyses indicated that time on task and number of attempted problems did not differ between conditions.[1]

3.2.3. Conditions compared on improvement in math skills and self-belief

Table 4 shows pretest and posttest scores on the math performance test and all self-belief measures, by condition. Difference scores were calculated for all measures. A MANOVA with condition as independent variable and difference scores as dependent variables showed a significant intercept, $F(5, 155) = 15.28$, $p < 0.001$, indicating that the difference scores differed from zero. Univariate tests showed that the difference between pretest and posttest was indeed statistically significant for math skills, Self-efficacy score (GMQ), and Perceived math competence, but not for the SMSMSS, and the Self-concept math score (SDQ I) meaning that children significantly improved their math skills and that the change in reported self-belief varied from no change to positive change. The effect of condition was not significant, $F(15, 471) = 1.13$, $p = 0.326$.

3.2.4. Conclusion

The results suggested that math practice was highest in the condition with medium success rate and in the choice condition, but the result was not confirmed with a non-parametric test. It is important to consider the power of Study 2. Table 3 showed that differences between conditions in math practice (i.e., number of attempted problems and time on task) were small, whereas standard deviations of these measures were relatively large. The size of the present sample was appropriate for detecting a medium effect size but a small effect size may have been more likely considering these standard deviations. Hence, the

[1] Two Kruskal-Wallis tests were performed with condition (4 levels: Choice, High, Medium, Low) as independent variable and number of attempted problems and time on task as dependent variable, respectively. Conditions did not differ in number of attempted problems (although the p-value was at trend level), $\chi2(3) = 7.29$, $p = 0.063$, and time on task in Math Garden, $\chi2(3) = 3.60$, $p = 0.308$.

Table 3
Math practice in Math Garden by condition.

Measure		Condition			
		Choice ($N = 48$)[a]	High success rate (N = 48)	Medium success rate (N = 48)	Low success rate ($N = 47$)
Attempted problems	Basic	455 (526)	435 (219)	482 (310)	339 (172)
	Bonus	356 (561)	359 (287)	375 (398)	300 (281)
Time on task (hrs)	Basic	0.96 (1.0)	0.73 (0.4)	1.04 (0.7)	0.82 (0.4)
	Bonus	0.78 (1.2)	0.67 (0.6)	0.83 (0.9)	0.69 (0.7)
p(High success rate)	Basic + Bonus	0.18 (0.22)	1 (0)	0 (0)	0 (0)
p(Medium success rate)	Basic + Bonus	0.35 (0.31)	0 (0)	1 (0)	0 (0)
p(Low success rate)	Basic + Bonus	0.47 (0.34)	0 (0)	0 (0)	1 (0)
Success rate		0.76 (0.07)	0.84 (0.04)	0.77 (0.04)	0.71 (0.05)

Note. Standard deviations between brackets.
[a] Data from one child were omitted in calculating the proportions of selected difficulty levels because this child never changed the difficulty level.

finding that the type of condition did not influence the motivation to practice math may just as well be attributed to a lack of power.

Considering the change in math skills, math skills of children who were allowed to select the difficulty of math problems themselves did not improve more than those of children who practiced with problems on a fixed difficulty level. Also, children in conditions with different fixed success rates improved comparably.

Results indicated that children as a group did not significantly improve their self-belief, with no differences between conditions. Table 4 shows that differences between conditions in self-belief measures were small, and that self-belief measures had large standard deviations. Hence, a conclusion on the effect of condition on self-belief is undetermined. It remains unclear whether the manipulations of (the choice of) success rate either did not affect self-belief or that the effect was too small to be detected in a sample of the current size.

4. Discussion

Use and benefit of an implementation of RSAT in a math practice application (Math Garden; Klinkenberg et al., 2011) were investigated. Study 1 concerned children's use of RSAT, the sample being the primary school children among the regular users of the application. RSAT allowed the choice for a high, medium, and low success rate and the relation between choices and children's gender, age, and ability were investigated. Study 2 concerned possible benefits for math practice, math performance, and self-belief, which were studied in a separate sample, recruited for the study. Study 2 used an in vivo experimental design where children from grades 3–6 practiced with Math Garden, either in a choice condition, where they could indicate their preferred success rate, or in a fixed-rate condition, where they played at a high (aimed success rate of 90%), medium (75%), or low success rate (60%).

Study 1 demonstrated that primary school children proved to be able to use the option of self-selecting the success rate. Results indicate that boys tend to prefer the low success rate more than girls do and that children with higher ability and in higher grades show higher preference for the low success rate than children with lower ability and children in lower grades. The observed gender difference is in accordance with the finding that girls evaluate their academic competences lower than boys do (Pomerantz et al., 2002). Pomerantz et al. (2002) suggest that girls are more anxious about encountering failures. Choosing the high success rate allows them to encounter less negative feedback. The observed positive relation between ability and preferred success rate is consistent with findings by Hontangas et al. (2000) and Johnson et al. (1991) but contrary to Revuelta's (2004) findings. An important advantage of the application under study is that the success rates are similar for all individuals, independent of ability. Success rates in other studies are based on the group average, which makes it difficult to decide whether a negative relation between ability and preferred success rate reflects matching the difficulty level to the student's ability or to a growing preference for *relatively* hard problems with increasing ability.

A shortcoming of Study 1 is that in Math Garden, the value of the coins earned when solving a problem correctly varies between conditions: the value of the coins is highest in the condition with low success rate and lowest in the condition with high success rate. Hence, it remains uncertain whether children select a success rate in order to manipulate the rate of success or to manipulate the value of the coins. Also, although an advantage of using computer adaptive practice is that it allows for using only three success rates, whereas studies on SAT commonly use more (Rocklin & O'Donnell, 1987), the relationships with gender, grade and ability might be sorted out more specifically using more than three success rates. Another limitation of Study 1 is the lack of control during data collection. Children play independently, at school or at home. We cannot control for environmental influences like noise in the direct environment, presence of family members or other students. However, it is unclear how such biases could explain

Table 4
Pretest and posttest scores on math performance test and self-belief measures by condition.

Measure		Condition			
		Choice	High success rate	Medium success rate	Low success rate
Math test (TTA)	Pre	64.7 (16.0)	63.7 (17.5)	64.8(16.8)	62.6 (14.5)
	Post	67.0 (16.3)	67.3 (17.9)	67.7 (16.7)	66.6 (14.3)
Self-efficacy (GMQ)	Pre	21.5 (3.1)	22.7 (4.0)	21.6 (3.2)	21.8 (4.1)
	Post	22.5 (3.5)	22.9 (4.0)	21.8 (3.4)	22.3 (4.1)
Self-concept math (SDQ I)	Pre	26.7 (7.0)	26.6 (8.1)	25.9 (7.1)	26.1 (7.5)
	Post	26.1 (6.6)	26.9 (7.5)	23.9 (7.6)	26.1 (7.6)
SMSMSS	Pre	105.4 (18.8)	108.1 (22.6)	100.9 (20.4)	102.6 (22.9)
	Post	103.2 (22.2)	106.5 (20.3)	98.0 (19.5)	102.8 (22.0)
Perceived math competence	Pre	17.4 (4.1)	16.9 (4.2)	16.8 (4.5)	18.4 (3.8)
	Post	18.1 (4.0)	18.0 (4.1)	17.2 (4.8)	18.5 (4.6)

Note. Standard deviations between brackets. TTA = Tempotoets Automatization; score shows the total of correctly solved addition and subtraction problems, range 0–100. GMQ = Global Motivation Questionnaire. SDQ I = Self-Description Questionnaire. SMSMSS = Sources of Middle School Mathematics Self -Efficacy Scale.

our effects. Given the robustness of the effects we conclude that the noise related to the data acquisition method is apparently compensated by the large amount of data and ecological validity of the data. Large data sets in natural environments can only be obtained in situations with limited control.

Study 2 followed up on the results of Study 1 and was directed to the question whether RSAT is beneficial for the motivation to practice math, improvement of math skills and self-beliefs concerning math. Contrary to expectations, children in all conditions showed a similar motivation to practice math in Math Garden. The present sample size was sufficient to detect a medium effect size. However, the indicators of the motivation to practice, number of attempted problems and time on task, showed high variance, making a small effect size more likely than a medium one. It remains unclear whether there was indeed no effect of choosing a success rate or that the power of Study 2 was too low to reject the null hypothesis.

An effect of choice on the motivation to practice is expected given the central role for autonomy in Self-determination theory (Ryan & Deci, 2000). Choice for a success rate truly changes the content of the task and relates to children's need for competence (Ryan & Deci, 2000). One may however argue that choice for a success rate does not change the type of problems (see also Reeve et al., 2003). It does change the difficulty of the problems but the problems are always math problems, no matter the choice. Also, the math application is applied in a practice situation, which is a low-stakes situation. In high-stakes situations, emotions are stronger and the need to adjust the environment accordingly may be more urgent than in low-stakes situations (see also Flowerday, Schraw, & Stevens, 2004). Next, practicing at the individual ability level may already increase motivation and reduce the need to adapt the success rate any further. Repeating Study 2 in a larger sample may shed light on these issues.

Children improved their math skills but the degree of improvement was independent of the possibility to choose a success rate and of the level of success-rate. The similar improvement in math skills across conditions with fixed, but various success rates was unexpected. Jansen et al. (2013) also used three conditions with fixed success rates and observed a positive relation between success rate, the motivation to practice, and improvement in math skills. The remaining condition in that study was a control group of children who did not practice in Math Garden, whereas the present study included a group of children who were allowed to choose the success rate. Another explanation for the absence of differences between the conditions with fixed success rates is that the true success rates in these conditions differed less than expected. Perhaps differences in success rates were not large enough to sort any effect on motivation to practice or on math skills.

Children did not improve their self-belief in Study 2, in none of the conditions. Perhaps self-belief concepts are rather stable and changing self-belief in a computer training of only 7.5 weeks is rather optimistic. Although periods of this length are used to decide whether interventions are effective for treating learning problems, this period may be too short to change self-beliefs and the period of intervention may be extended in future research. Otherwise, the questionnaires applied may not have been specific enough (Pajares, 1996). Given the large standard deviations of the scores on the questionnaires, this part of Study 2 has probably also suffered from low power.

The sample size of Study 2 was limited for the large number of conditions. Although two conditions (choice versus no choice) would have been sufficient to investigate whether choice would have a beneficial effect, the success rate in the no choice condition was hard to define in advance. Choice may be beneficial in comparison to practicing hard problems but not in comparison to practicing easy problems, or the other way around. The many conditions as well as the large variance in some outcome measures lowered the probability of finding significant effects.

A final explanation for the absence of significant results in Study 2 concerns the representativeness of the sample. Although scores on the math skills test are comparable to those in the normal population (De Vos, 2010) and relatively high self-belief scores are common for Dutch children (Lee, 2009), the sample shows a relatively high preference for the low success rate, higher than in the large and representative sample of Study 1. Hence, the children in Study 2 may have been attracted to the higher value of the coins in the low success rate or to the difficulty of the problems. Group dynamics may have caused children to select the lowest success rate, which may however not always fit their individual preferences. Perhaps children in Study 2 did not follow their individual preference for a success rate.

Combined, the results of Studies 1 and 2 are important for the classroom and show the potential of computer-assisted learning. The results of Study 1 are based on exceptionally large numbers of children (>40,000), and a wide age range (grades 1–6) and indicate that implementation of RSAT in a math practice application seems to fulfill a desire to adapt the balance between challenge and success to the individual need. Every possibility to increase the attractiveness of practicing math is a benefit because math is essential in both education and everyday life, but is still not very much appreciated by (all) children (Dowker, 2005).

Acknowledgements

We would like to thank Nadine Bürmann for her help in setting up and collecting data for Study 2. The work of the second author and third author is funded by the Netherlands Organisation for Scientific Research (NWO), Grant Nr 406-11-136, and CI1-12-S037, respectively.

References

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology*, *130*(2), 224–237.

Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, *14*(2), 243–248.

Ashcraft, M. H., Krause, J. A., & Hopko, D. R. (2007). Is math anxiety a mathematical learning disability. In D. B. Berch, & M. M. Mazzocco (Eds.), *Why is math so hard for some children* (pp. 329–348). Baltimore, MD: Brookes.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Cross, C. P., Copping, L. T., & Campbell, A. (2011). Sex differences in impulsivity: A meta-analysis. *Psychological Bulletin*, *137*, 97–130. http://dx.doi.org/10.1037/a0021591.

De Vos, T. (2010). *Manual tempotoets automatiseren.* Amsterdam: Boom Testuitgevers.

Dowker, A. (2005). *Individual differences in arithmetic: Implications for psychology, neuroscience, and education.* Hove: Psychology Press.

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. http://dx.doi.org/10.1177/0146621606288890.

Elo, A. (1978). *The rating of chessplayers.* Arco: Past and Present.

Erturan, S., & Jansen, B. R. J. (2015). Boys' and girls' emotional experience of math: The same scores but a different relation with math performance. *European Journal of Psychology of Education*, *4*, 421–435.

Flowerday, T., Schraw, G., & Stevens, J. (2004). The role of choice and interest in reader engagement. *The Journal of Experimental Education*, *72*(2), 93–114. http://dx.doi.org/10.3200/JEXE.72.2.93-114.

Harter, S. (1982). The perceived competence scale for children. *Child Development*, *53*, 87–97.

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, *21*, 33–46.

Hontangas, P., Olea, J., Ponsoda, V., Revuelta, J., & Wise, S. L. (2004). Assisted self-adaptive testing: A comparative study. *European Journal of Psychological Assessment*, *20*, 2–9. http://dx.doi.org/10.1027/1015-5759.20.1.2.

Hontangas, P., Ponsoda, V., Olea, J., & Wise, S. L. (2000). The choice of item difficulty in self-adapted testing. *European Journal of Psychological Assessment*, *16*, 3–12. http://dx.doi.org/10.1027/1015-5759.16.1.3.

Jansen, B. R. J., Louwerse, J., Straatemeier, M., Van der Ven, S. H. G., Klinkenberg, S., & der Maas, V. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, *24*, 190–197. http://dx.doi.org/10.1016/j.lindif.2012.12.014.

Johnson, P. L., Roos, L. L., Wise, S. L., & Plake, B. S. (1991). Correlates of examinee item choice behavior in self-adapted testing. *Mid-Western Educational Researcher*, *4*, 25–28.

Katz, I., & Assor, A. (2007). When choice motivates and when it does not. *Educational Psychology Review*, *19*, 429–442.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education*, *57*(2), 1813–1824. http://dx.doi.org/10.1016/j.compedu.2011.02.003.

Krinzinger, H., Kaufmann, L., & Willmes, K. (2009). Math anxiety and math ability in early primary school years. *Journal of Psycho educational Assessment*, *27*, 206–225. http://dx.doi.org/10.1177/0734282908330583.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences*, *19*(3), 355–365. http://dx.doi.org/10.1016/j.lindif.2008.10.009.

Li, H., Yuan, J., & Lin, C. (2008). The neural mechanism underlying the female advantage in identifying negative emotions: An event-related potential study. *NeuroImage*, *40*, 1921–1929. http://dx.doi.org/10.1016/j.neuroimage.2008.01.033.

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education*, *30*, 520–540.

Maris, G., & Van der Maas, H. L. J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633. http://dx.doi.org/10.1007/S11336-012-9288-Y.

Marsh, H. W. (1988). *Self-description questionnaire: A theoretical and empirical basic for the measurement of multiple dimensions of preadolescent self-concept: A test manual and a research monograph.* San Antonio, TX: The Psychological Corporation.

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *British Journal of Educational Psychology*, *81*, 59–77.

Meyer, D. K., Turner, J. C., & Spencer, C. A. (1997). Challenge in a mathematics classroom: Students' motivation and strategies in project-based learning. *The Elementary School Journal*, *97*(5), 501–521.

Moeller, S. K., & Robinson, M. D. (2010). Sex differences in implicit punishment sensitivity: Evidence from two cognitive paradigms. *Personality and Individual Differences*, *48*, 283–287. http://dx.doi.org/10.1016/j.paid.2009.10.017.

Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competency, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education*, *7*(2), 133–144.

Ortner, T. M., & Caspers, J. (2010). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, *27*(3), 157–163.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578.

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computered adaptive tests: A meta-analysis. *Journal of Educational Measurement*, *38*, 235–247. http://dx.doi.org/10.1111/j.1745-3984.2001.tb01125.x.

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology*, *94*(2), 396–404. http://dx.doi.org/10.1037//0022-0663.94.2.396.

Ponsoda, V., Wise, S. L., Olea, J., & Revuelta, J. (1997). An investigation of self-adapted testing in a Spanish high school population. *Educational and Psychological Measurement*, *57*(2), 210–221. http://dx.doi.org/10.1177/0013164497057002002.

Prast, E. J., & Van de Weijer-Bergsma, E. (2013). *Handleiding globale motivatievragenlijst, version april 2013.* The Netherlands: Universiteit Utrecht (unpublished internal document).

Reeve, J., Nix, G., & Hamm, D. (2003). Testing models of the experience of self-determination in intrinsic motivation and the conundrum of choice. *Journal of Educational Psychology*, *95*(2), 375–392.

Revuelta, J. (2004). Estimating ability and item-selection strategy in self-adaptive testing: A latent class approach. *Journal of Educational and Behavioral Statistics*, *29*, 379–396. http://dx.doi.org/10.3102/10769986029004379.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology*, *79*, 315–319. http://dx.doi.org/10.1037//0022-0663.79.3.315.

Roos, L. L., Wise, S. L., & Finney, S. J. (1998). *Comparing restricted and unrestricted self-adapted testing as alternatives to computerized adaptive testing. Paper presented at the annual meeting of the national council on measurement in education.* CA: San Diego.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

Schirmer, A., Zysset, S., Kotz, S. A., & von Cramon, D. Y. (2004). Gender differences in the activation of inferior frontal cortex during emotional speech perception. *NeuroImage*, *21*(3), 1114–1123. http://dx.doi.org/10.1016/j.neuroimage.2003.10.048.

Shermis, M. D., Mzumara, H. R., & Bublitz, S. T. (2001). On test and computer anxiety: Test performance under CAT and SAT conditions. *Journal of Educational Computing Research*, *24*, 57–75. http://dx.doi.org/10.2190/4809-38LD-EEUF-6GG7.

Simons, J., & Fisette, P. (2001). Het zelfconcept bij kinderen uitgaande van een multidimensioneel hiërarchisch model. In J. Simons (Ed.), *Actuele themata uit de psychomotorische therapie, jaarboek 1* (pp. 79–98). Leuven, Belgium: Acco.

Usher, E. L., & Pajares, F. (2006). Sources of academic and self-regulatory efficacy beliefs of entering middle school students. *Contemporary Educational Psychology*, *31*, 125–141.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, *78*(4), 751–796.

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: A validation study. *Contemporary Educational Psychology*, *34*, 89–101.

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

Van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2012). The development of executive functions and early mathematics: A dynamic relationship. *British Journal of Educational Psychology*, *82*, 100–119. http://dx.doi.org/10.1111/j.2044-8279.2011.02035.x.

Van Duijvenvoorde, A. C. K., Zanolie, K., Rombouts, S. A. R. B., Raijmakers, M. E. J., & Crone, E. A. (2008). Evaluating the negative of valuing the positive? Neural mechanisms supporting feedback-based learning across development. *Journal of Neuroscience*, *28*, 9495–9503. http://dx.doi.org/10.1523/JNEUROSCI.1485-08.2008.

Veerman, J. W., Straathof, M. A. E., Treffers, P. D. A., Van den Bergh, B. R. H., & Ten Brink, L. T. (2000). *Dutch manual for perceived competence scale for children.* Amsterdam: Pearson.

Vispoel, W. P., & Coffman, D. D. (1994). Computerized-adaptive and self-adapted music-listening tests: Psychometric features and motivational benefits. *Applied Measurement in Education*, *7*(1), 25–51. http://dx.doi.org/10.1207/s15324818ame0701_4.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adaptive testing. *Applied Measurement in Education*, *7*(1), 53–79. http://dx.doi.org/10.1207/s15324818ame0701_5.

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

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473–492.

Weller, J. A., Levin, I. P., & Bechara, A. (2009). Do individual differences in Iowa gambling task performance predict adaptive decision making for risky gains and losses? *Journal of Clinical and Experimental Neuropsychology*, *32*, 141–150. http://dx.doi.org/10.1080/13803390902881926.

Williams, T., & Williams, K. (2010). Self-efficacy and performance in mathematics: Reciprocal determinism in 33 nations. *Journal of Educational Psychology*, *102*(2), 453–466.

Wise, S. L., Kingsbury, G. G., & Houser, R. L. (1993). *An investigation of restricted self-adapted testing. Paper presented at the annual meeting of the national council on measurement in education.* GA: Atlanta.

Yuan, J., Luo, Y., Yan, J. H., Meng, X., Yu, F., & Li, H. (2009). Neural correlates of the females' susceptibility of negative emotions: An insight into gender-related prevalence of affective disturbances. *Human Brain Mapping*, *30*, 3676–3686. http://dx.doi.org/10.1002/hbm.20796.