Governing games

Adaptive game selection in the Math Garden

Matthieu Brinkhuis[1,*], *Wessel* Cordes, and *Abe* Hofman[2,3]

[1]Utrecht University, Information and Computing Sciences, Utrecht, 3584 CC, The Netherlands
[2]University of Amsterdam, Psychological Methods, Amsterdam, 1018 WS, The Netherlands
[3]Prowise Learn, Amsterdam, 1011 VL, The Netherlands

Abstract. Utilizing online digital educational content has become the norm when teaching young students. A variety of adaptive educational practice systems is readily available and allows students to practice various domains, on a preferred difficulty and pace. However, due to the intensification of the teaching profession and the possibilities of practicing from home, students might be left unsupervised, and as a result do not practice domains that are most important. This study proposes a solution to *govern* these students, i.e., provide computerized data driven supervision that guides students in practicing domains most important with no intervention of a teacher.
Through an experiment involving 13 578 participants, a new governing method was tested and found to have positive effects on both engagement and learning, with almost no changes to the visual interface needed. Governing seems a promising technique in general, and was effectively tested and introduced in Math Garden.

1 Introduction

Utilizing online digital educational content has become the norm when teaching young students. For example, there are many adaptive educational practice systems allow students to practice their arithmetic abilities in various domains on a preferred difficulty and pace, such as ALEKS, Knewton and Math Garden [1–3].

However, due to the intensification of the teaching profession and the possibilities of practicing from home, students are often left unsupervised, and as a result do not practice domains that are most important. Therefore, this research proposes technology to govern their learning experience and to increase student engagement and learning. Governing is defined as computerized data driven supervision that guides students in practicing domains most important without intervention of a teacher. Three governing principles are laid out in this paper, and applied in an experiment in the online practice environment Math Garden [4].

Math Garden originated in 2007 as a tool to study the dynamics of cognitive development in children, specifically the development of mathematical knowledge and abilities. The Math Garden developed into a large-scale online learning system, and more than a decade worth of data has been analyzed [3]. See figure 1 for a screenshot of the landing page. Fundamentally, Math Garden is a computerized adaptive practice (CAP) system aimed towards primary

*e-mail: m.j.s.brinkhuis@uu.nl

Figure 1. A screenshot of the landing page of the Math Garden, showing a selection of 8 games, representing 8 practice domains. Note that 4 plants appear withered and have watering cans in place, representing a naive governing method.

education. The adaptive element is a modification of the Item Response Theory (IRT) approach and is based on the Elo [5–7] rating system combined with response times [8, 9]. The model estimates the ability of the student and the difficulty of the item (i.e. question [10]). The estimation is updated after every answered item, allowing for on the fly calibration. As of 2018, more than 452 000 K-12 children and 5 000 Dutch primary schools used Math Garden to practice arithmetic, completing roughly a million items every day, with a total of 853 300 000 item responses generated.

Math Garden currently consists of 23 games, referred to as domains, and new domains are released regularly, hence students may be overwhelmed with choice. Not all 23 domains are immediately playable, since some domains are unlocked after reaching a certain level in another domains or depend on the grade of the student. Still, most students have access to over 12 games. These games are divided in a *base* and *bonus* garden. Base garden domains relate to basic skills students need to learn, whereas bonus garden domains, though educational, do not directly relate to basic skills and are sometimes considered more enjoyable or more interesting by users. While there is some form of control from the system over what domains are available, see figure 1 and section 2.4, there is no supervision over what domains should be practiced. In Math Garden this has led to unwanted behavior [11], where exploratory analyses show that quite some older students still practice preschool domains, and that quite some students only play a few domains. This information indicates there is a strong need for guidance from the system in Math Garden, and has been a motivation to research methods for governing games.

2 Methods

Game recommendations in educational games with many domains to choose from are clearly beneficial, especially regarding the relations between domains [12], which can be used to suggest what to practice. We define governing [13] as computerized, data driven supervision

in an educational practice system, with three specific goals: 1) an individualized domain selection mechanism, 2) guide students in selecting best domains to practice and 3) increase most important abilities. We present a naive approach and present a simple updated governing method with an application in the Math Garden, which is validated through an online experiment. To meet these goals, three governing principles are defined: governing perspectives, strategies and visualizations, which are elaborated upon below.

2.1 Governing perspectives

Governing can be driven by different perspectives. First, governing could be driven from education, e.g., following educational guidelines such as a (national) curriculum. In the Netherlands the national institute of curricular development has clear guidelines concerning grade specific math and arithmetic development in students [14]. A second perspective is to look at governing from a practice system specific perspective, which can be data driven. For example, expected learning gains can be quantified and translated to what domains to practice next. A third perspective is peer oriented. Here, students are compared to peer reference groups, either national grade reference distributions, or perhaps class specific reference distributions per domain. A related perspective concerns the reliable estimation of the ability of new players in the system, also know the cold-start problem, see for example [15]. A simple approach would be to drive the governing such that domains with a small number of (recent) observations are more often selected - the used approach should be related to how the cold-start problem are alleviated. Clearly, other approaches might be thought of, and the mentioned approaches can be used in combination.

2.2 Governing strategies

Following [16], we can define the following governing strategies.

A naive strategy does not learn and uses existing information to solve a problem. For example, one can compare the student's knowledge to a static target knowledge level or baseline, obtained from one of the perspectives mentioned before. A naive strategy might select domains furthers below the baseline.

An expert strategy keeps learning, for example predicting student domain knowledge based on knowledge in other domains [12, 17], such as a correlation matrix of all domains. Predictions based on such a model can be used for domain selection to maximize learning gain.

An experiment driven strategy is an iterative testing strategy and allows for evaluating learning interventions and reveal pattern and side-effects [18]. Online experiments, or A/B tests, determine what is needed to attain a certain knowledge level or domain rating.

2.3 Governing visualizations

The last governing principle relates to the visualization of the selected domains. Governing perspectives and strategies determine what domains should be practiced, yet how should the selection be presented? Two axes are determined for governing visualizations: they can be flexible or strict in several aspects, and rewarding or punishing.

Flexible visualization offers selected domains as, possibly mild, recommendations. The option is open to adhere to the recommendation. Forcing these choices would be a strict approach. Stimulating preferred domain choices can be approached in many ways and often attempt to stimulate intrinsic motivation, for example through highlighting domains [19] or

using gamification elements [20, 21]. For strict visualizations, a decrease in task engagement can be expected [22].

A rewarding visualization uses extrinsic motivation, through rewarding students for adhering to the preferred domains [23, 24]. A punishing visualization is the opposite, punishes students for not engaging in an activity, like diminishing levels, points, progress, etc.

2.4 Designing a new governing method

Math Garden already has some basic form of governing in place. The main principle this basic governor functions on, is recent practice. As shown in figure 1, plants representing domains in the base garden start to wither when a domain is not practiced for over 9 days, and watering cans appear. Though base garden domains are always accessible, the presence of a watering can limits access to all bonus garden games. Watering cans are erased once a student finishes a practicing session in that specific domain. Hence, this governing method carries out a punishing, but flexible governing visualization. The new governor is designed to follow the governing principles laid out before.

Concerning governing perspectives, it combines all three approaches. First, it is domain oriented in that it defines relevant and non-relevant domains for the current grade level following a national curriculum [14]. The domains of the national curriculum are mapped to the Math Garden domains for this purpose, as presented in table 1. In addition, using data from the practice system peer group reference distributions are created for each grade. These reference distributions allow to quantify the distance to one's peers, and order domains accordingly.

ID	Domain	1	2	3	4	5	6	7	8
64	Shape and color	■	■						
7	Counting	■	■	■					
41	Numbers		■	■	■	■			
1	Addition				■	■	■	■	
2	Subtraction				■	■	■	■	
9	Clocks				■	■	■		
3	Multiplication					■	■	■	■
4	Division					■	■	■	■
...	...								

Table 1. A selection of the Math Garden domain IDs mapped to the Dutch national curriculum domains for grades 1-8.

Concerning government strategies, a naive strategy is chosen where calculation on reference distributions are made beforehand and evaluated using an A/B test, which is elaborated upon in section 2.5.

As visualization, a quite basic approach was taken where the interface of the Math Garden with the watering cans suggesting domains to practice, figure 1 was maintained. However, the flexibility/strictness approach was changed. Where watering cans would in the old approach appear after nine days of not practicing a domain, possibly leading to all base domains in need of practice after a short break, a maximum of 3 watering cans was foreseen in the new governor to provide more flexibility. The rewarding/punishing visualization aspects were kept the same during the experiment.

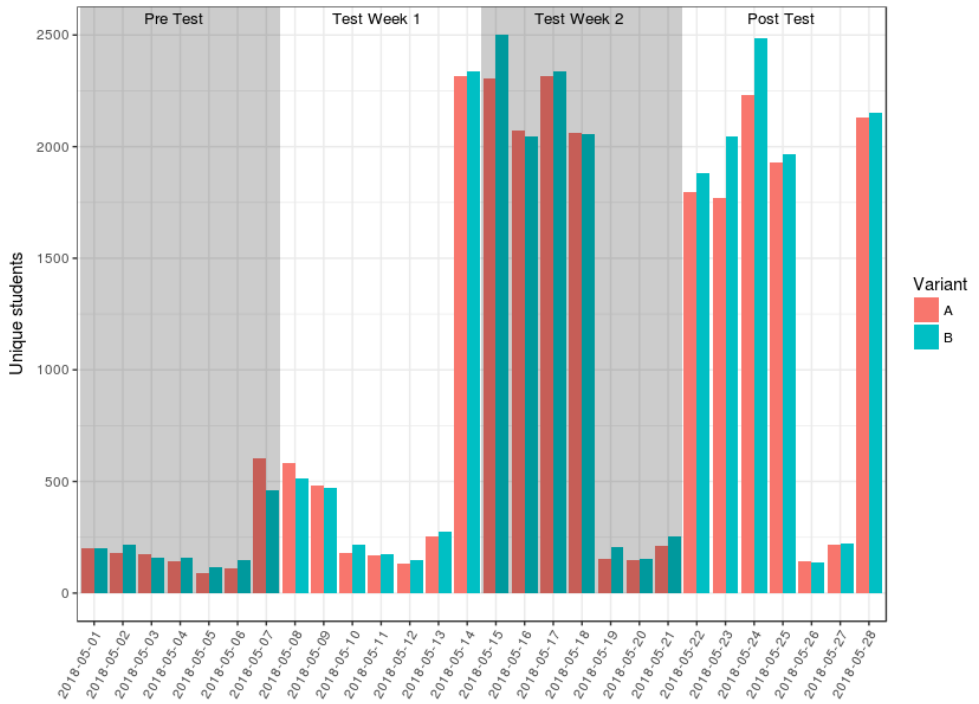

Figure 2. Daily number of responses in the Math Garden for the two testing variants A (control) and B (experiment). Several national holidays reduced the number of responses on these days.

2.5 Online experiment

The governing method was evaluated in an A/B test [25, 26] running for two full weeks [27]. Students that were or had been actively practicing in the Math Garden were randomly assigned to the control variant A (default system) or treated with the governing method (variant B). The visual impact on participants is minimal, the most noticeable change being that students have no more than three watering cans present in their garden. The two evaluation criteria of the experiment were engagement and learning.

3 Results

The solution governing method was found to have positive effects on both engagement and learning, with almost no changes to the visual interface needed.

First, some general numbers on the A/B test are presented, followed by a section on the effects on engagements and a section on the effects on student learning.

A total of 13 578 students participated in the experiment. 6 785 students were in control variant A (default system) and 6 793 were treated with the solution governing method (variant B). The distribution of students over the different days of the experiment can be found in figure 2.

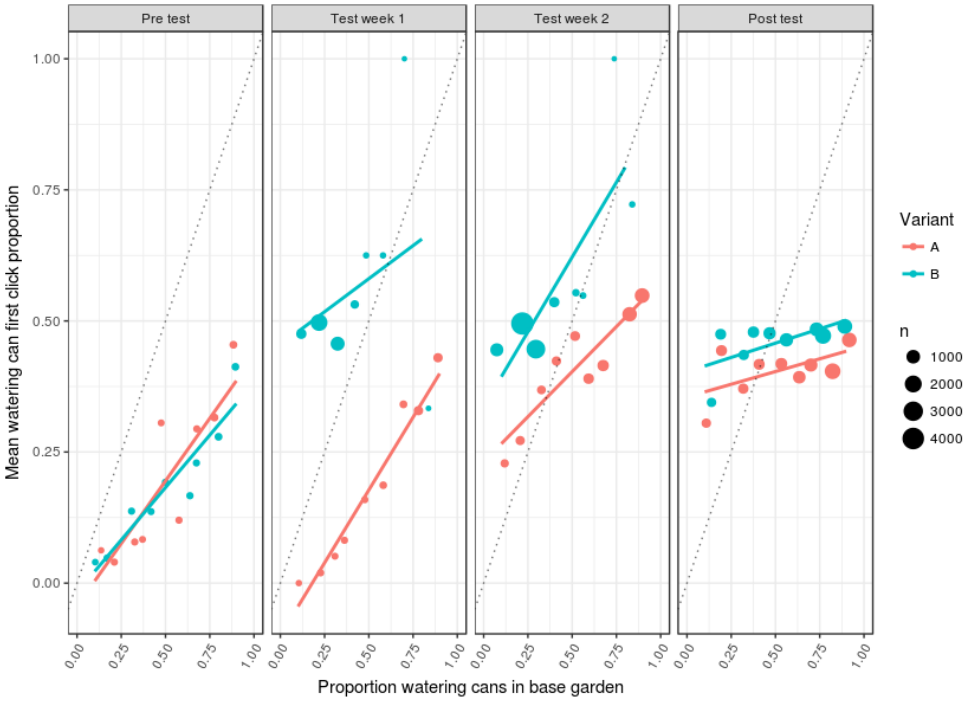

Figure 3. First click rates per A (control) and B (experimental) test condition, per test week. Points above the diagonals (dotted) indicate an average attraction to the watering can domains, points below the diagonal an average avoidance of watering can domains.

3.1 Engagement

Engagement has been measured using two constructs. First click rates, what domain the student click on first, is relevant because the governor should stimulate what domain to practice first. Finishing rates, whether students finish the domain they choose, is relevant because just clicking on a domain and exiting is an indication governing is not working properly.

Figure 3 shows the first click rates for each of the conditions. Since the number of watering cans can be different in both conditions, and the number of active games can differ between students, it was decided to use the proportion of watering cans, represented on the horizontal axis. On the vertical axis, we see the proportion of times a first click is on a watering can. Hence, if all points in this figure would be on the diagonal, domains with watering cans are as attractive on average as domains without watering cans. Points above the diagonal indicate watering can domains are more attractive, and points below the diagonal indicate watering can domains are less attractive. Hence, it should be noted that in the pretest period children are actively avoiding the watering cans. This can be explained by the fact that the games with watering cans are often the more difficult and less attractive games for the student.

A large positive effect on engagement is shown during the test weeks, where in the experimental condition the watering can domains are more attractive. The differences in the proportion of watering in variant B during the test weeks are exclusively related to differences in the number of active games.

Another approach to quantify engagement is to look at finishing rates. Figure 4 shows the difference between finishing rates of the control (A) and treatment (B) condition, where

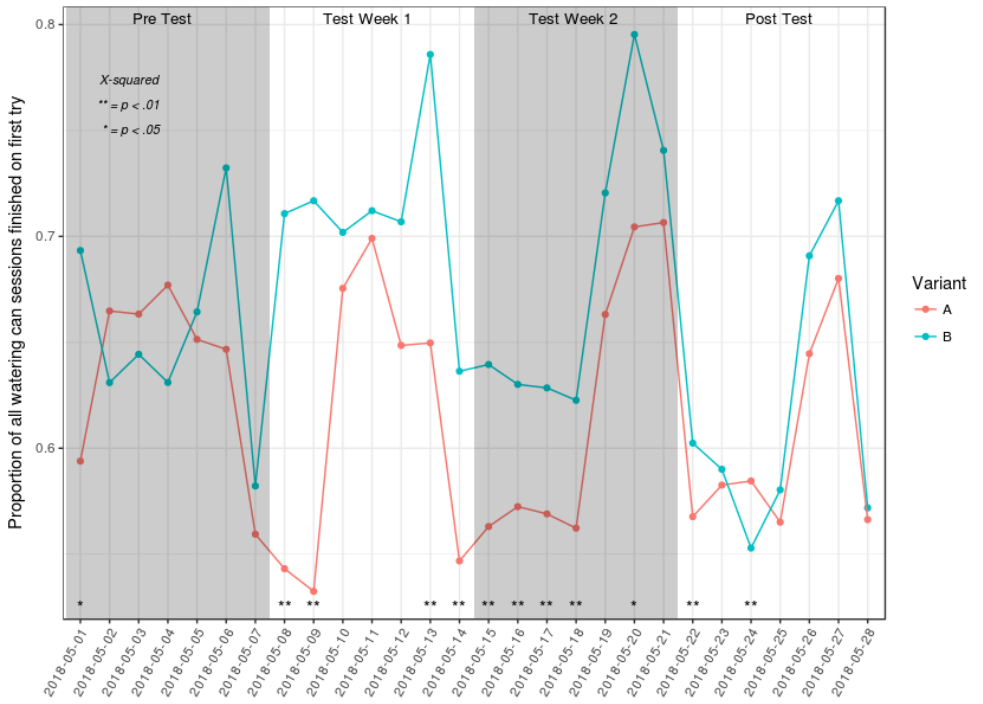

Figure 4. Differences between finishing rates of the control (A) and treatment (B) condition over the length of the entire A/B test.

clearly finishing rates of the treatment condition are higher during the test weeks. A minor post-test effect can be observed too.

3.2 Learning

Next to engagement, possible effects on learning rates are regarded important. Learning can be quantified in multiple ways, here a simple approach has been taken where learning is quantified as 1) number of completed items in (relevant) base domains, 2) number of domains on which the students score below their peer grade reference distributions and 3) differences in estimated ability in the rating system.

The number of relevant base domains in which no items were practiced was considered problematic for participants in the learning environment, partly because for these domains reliable ability estimates are missing. The number of domains with the number of items practiced below 30 was counted for every participant before and after the experiment. Clearly, since students practice in both experimental conditions, we expect the number of domains with fewer than 30 practiced items to decrease. Table 3.2 shows that on average, the number of non-practiced domains drops .29 for the treatment conditions verses .23 for the base condition, a significant difference. Similarly to counting the reduction in number of domains with few items played, we can count the number of domains that have a lower score than their peer reference group. Again we expect this number to reduce due to practice, yet as seen in table 3.2, this number drops significantly faster for the new governing method.

Finally, we can evaluate the effect on ability by using the Math Gardens internal ratings, denoted as θ. For each domain, the average daily change in rating $\Delta\theta$ is plotted in figure 5. Here

Variable	Count	Mean
Relevant base domains		
Variant A	6 703	0.13
Variant B	6 713	0.21**
All base domains		
Variant A	6 708	0.23
Variant B	6 719	0.29**

Table 2. Average decrease in number of domains with fewer than 30 practiced items. The decrease is higher for the new governing method, indicating on average fewer domains have close to no items practiced after the experiment.

Variable	Count	Mean
Relevant base domains		
Variant A	6 703	0.09
Variant B	6 713	0.16**
All base domains		
Variant A	6 708	0.17
Variant B	6 719	0.21**

Table 3. Average decrease of domains with below peer score baseline. The decrease is higher for the new governing method, indicating on average fewer domains have scores below a peer score baseline.

it can be seen that even for this rather short experiment, there seems a consistent positive effect on daily ability increase. Fitting a linear mixed model shows that the rating development in variant B (treatment) compared to variant A (control) over the entire experiment is significantly higher with a 20% difference.

4 Discussion

This research explored governing and provides the first steps, knowledge and reasoning to introduce governing in educational practice systems. By analyzing data from a large scale A/B test, we found that the implemented governing procedure improved the engagement and learning rates of players in the Math Garden.

The presented results illustrate positive effects of the current implementation compared to the earlier implementation in the Math Garden, even though the intervention was not complex and the duration of the experiment rather short. However, this short duration of the experiment prevented to test long term learning effects. A long term experiment, including ability testing outside the learning environment, should also shed light on whether the learning effect can generalize and if these are related to other phenomena, such as cold-start problems.

The current governing implementation combines information on (1) reliability of the measurements (number of made items for each domain), (2) the ability on the domain compared to other domains, (3) the ability compared to peers and (4) some motivational aspects, for example allowing domains not be recycled each day. This motivational aspect prevents that children are too often confronted with their weakest domains, possibly lowering their motivation to play. How each of these sources of information are used in an optimal way, both

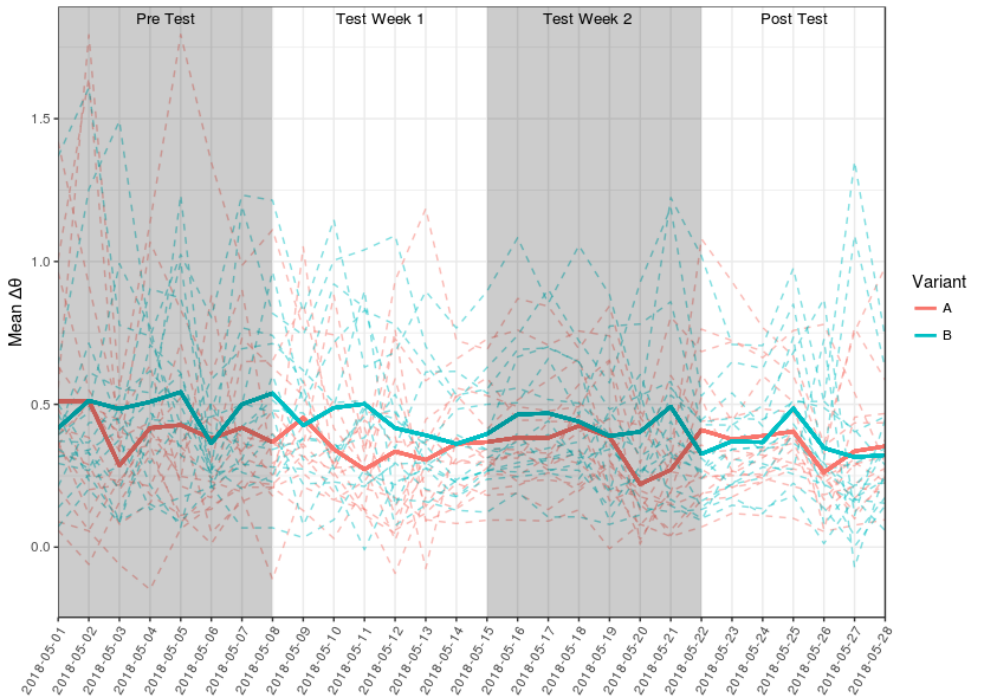

Figure 5. Mean daily $\Delta\theta$ per base garden domains (dotted), mean $\Delta\theta$ (solid). The treatment condition average seems consistently above the control condition average, indicating a faster development in ability on average.

individually and combined, should be further researched, for example by using an A/B testing framework.

To conclude, we presented first steps in exploring governing, and performing a randomized experiment in governing the behavior of children in an online learning environment. Based on the current research findings in Math Garden, an actual extension of the governing algorithm was implemented for all students, adding a priority ranking within the selected domains and an achievement bonus for playing a selected domain. With the further improvement of the algorithm and the visualization, taking engagement, motivational effects and the reliable measurement of abilities into account, we hope to gently guide the behavior of children to optimize their engagement and learning.

The authors declare no conflict of interest. The governing module was implemented in collaboration with Prowise Learn, owner of the Math Garden. Prowise had no role in the data analyses, the interpretation of the data, in the writing of the manuscript, or in the decision to publish the results. The anonymized data can be requested by contacting A.H. This work is based on the Master thesis of W.C. [28].

References

[1] J.C. Falmagne, E. Cosyn, J.P. Doignon, N. Thiéry, in *Formal concept analysis* (Springer, 2006), pp. 61–79

[2] K. Wilson, Z. Nichols, White paper, Knewton, New York, NY (2015), `http://learn.knewton.com/technical-white-paper`

[3] M.J.S. Brinkhuis, A.O. Savi, F. Coomans, A.D. Hofman, H.L.J. van der Maas, G. Maris, Journal of Learning Analytics **5**, 29 (2018)
[4] S. Klinkenberg, M. Straatemeier, H.L.J. van der Maas, Computers & Education **57**, 1813 (2011)
[5] A.E. Elo, *The rating of chess players, past and present* (B. T. Batsford, Ltd., London, 1978)
[6] M.J.S. Brinkhuis, G. Maris, Measurement and Research Department Reports 09-01, Cito, Arnhem (2009), https://www.researchgate.net/publication/242357963
[7] M.J.S. Brinkhuis, Ph.D. thesis, University of Amsterdam (2014), http://hdl.handle.net/11245/1.433219
[8] G. Maris, H.L.J. van der Maas, Psychometrika **77**, 615 (2012)
[9] F. Coomans, A.D. Hofman, M.J.S. Brinkhuis, H.L.J. van der Maas, G. Maris, PLOS ONE **11**, 1 (2016)
[10] M.J.S. Brinkhuis, M. Bakker, G. Maris, Journal of Educational Measurement **52**, 319 (2015)
[11] S. Sosnovsky, L. Müter, M. Valkenier, M. Brinkhuis, A. Hofman, in *Lifelong Technology-Enhanced Learning*, edited by V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, M. Scheffel (Springer, 2018), pp. 531–536, ISBN 978-3-319-98572-5
[12] A.D. Hofman, B.R.J. Jansen, S.M.M. de Mooij, C.E. Stevenson, H.L.J. van der Maas, Journal of Intelligence **6** (2018)
[13] F. Fukuyama, Governance **26**, 347 (2013)
[14] SLO, *TULE inhouden & activiteiten* (2018), http://tule.slo.nl/index.html
[15] J.Y. Park, S.H. Joo, F. Cornillie, H.L.J. van der Maas, W. Van den Noortgate, Behavior Research Methods **51**, 895 (2019)
[16] D. Gentner, A.L. Stevens, *Mental models* (Psychology Press, 2014)
[17] A. Hofman, R. Kievit, C. Stevenson, D. Molenaar, I. Visser, H.L.J. van der Maas (2018)
[18] A.O. Savi, J.J. Williams, G. Maris, H.L.J. van der Maas, *The role of A/B tests in the study of large-scale online learning* (2017)
[19] S. Singh, Management decision **44**, 783 (2006)
[20] S. Barab, M. Thomas, T. Dodge, R. Carteaux, H. Tuzun, Educational technology research and development **53**, 86 (2005)
[21] Z.H. Chen, C. Liao, H. Cheng, C. Yeh, T.W. Chan, Journal of Educational Technology & Society **15**, 317 (2012)
[22] R. Ryan, E. Deci, American psychologist **55**, 68 (2000)
[23] R. Vallerand, L. Pelletier, M. Blais, N. Briere, C. Senecal, E. Vallieres, Educational and psychological measurement **52**, 1003 (1992)
[24] J. Hamari, J. Koivisto, H. Sarsa, *Does gamification work?–a literature review of empirical studies on gamification*, in *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (IEEE, 2014), pp. 3025–3034
[25] A.O. Savi, H.L.J. van der Maas, G.K.J. Maris, Science **347**, 958 (2015)
[26] A.O. Savi, N.M. Ruijs, G.K.J. Maris, H.L.J. van der Maas, Computers & Education **119**, 84 (2018)
[27] R. Kohavi, R. Longbotham, in *Encyclopedia of Machine Learning and Data Mining* (Springer, 2017), pp. 922–929
[28] W. Cordes, mathesis (2018), https://dspace.library.uu.nl/handle/1874/368080