UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

# Learning as It Happens: A Decade of Analyzing and Shaping a Large-Scale Online Learning System

Brinkhuis, M.J.S.; Savi, O.A.; Hofman, A.D.; Coomans, F.; van der Maas, H.L.J.; Maris, G.

**DOI**
10.18608/jla.2018.52.3

**Publication date**
2018

**Document Version**
Final published version

**Published in**
Journal of Learning Analytics

**License**
CC BY-NC-ND

Link to publication

**Citation for published version (APA):**
Brinkhuis, M. J. S., Savi, O. A., Hofman, A. D., Coomans, F., van der Maas, H. L. J., & Maris, G. (2018). Learning as It Happens: A Decade of Analyzing and Shaping a Large-Scale Online Learning System. *Journal of Learning Analytics*, *5*(2), 29-46. https://doi.org/10.18608/jla.2018.52.3

**General rights**
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)*

Download date:09 Mar 2023

JOURNAL OF LEARNING ANALYTICS

Volume 5(2), 29–46. http://dx.doi.org/10.18608/jla.2018.52.3

# Learning As It Happens: A Decade of Analyzing and Shaping a Large-Scale Online Learning System

Matthieu J. S. Brinkhuis [1], Alexander O. Savi [2], Abe D. Hofman [3], Frederik Coomans [4], Han L. J. van der Maas [5], Gunter Maris [6]

**ABSTRACT**

With the advent of computers in education, and the ample availability of online learning and practice environments, enormous amounts of data on learning become available. The purpose of this paper is to present a decade of experience with analyzing and improving an online practice environment for math, which has thus far recorded over a billion responses. We present the methods we use to both steer and analyze this system in real-time, using scoring rules on accuracy and response times, a tailored rating system to provide both learners and items with current ability and difficulty ratings, and an adaptive engine that matches learners to items. Moreover, we explore the quality of fit by means of prediction accuracy and parallel item reliability. Limitations and pitfalls are discussed by diagnosing sources of misfit, like violations of unidimensionality and unforeseen dynamics. Finally, directions for development are discussed, including embedded learning analytics and a focus on online experimentation to evaluate both the system itself and the users' learning gains. Though many challenges remain open, we believe that large steps have been made in providing methods to efficiently manage and research educational big data from a massive online learning system.

**Notes for Practice**

- We analyzed an online adaptive practice environment for arithmetic, actively used by over 400,000 primary school children in the Netherlands.
- Adaptive practice is achieved by continuously tracking both student abilities and item difficulties, and matching students to items.
- A unidimensional adaptive algorithm, separately employed within each domain (e.g., multiplication), takes care of tracking abilities and difficulties.
- We show that the obtained unidimensional ability and difficulty estimates are, to a large extent, reliable and accurate.
- Moreover, we explore the many sources of misfit, or violations of the unidimensionality assumption, including differences in response processes (fast and slow responders) and response strategies (erroneous strategies that work for clusters of items).
- To advance the field of learning analytics, we call for *active* analytics such as exemplified in this paper. Learning analytics must actively help direct a student towards his or her educational objective by means of embedded analytics that not only analyze the student, but also shape their learning path (such as the discussed adaptive algorithm) and includes experiments that ensure changes to the system have the desired effect.

**Keywords**

Adaptive learning, educational games, exploring quality of fit, adaptive item selection, evaluation of CAL systems

**Submitted:** 14.10.18 — **Accepted:** 05.02.18 — **Published:** 05.08.18

*Corresponding author [1]Email: m.j.s.brinkhuis@uu.nl Address: Utrecht University, Department of Information and Computing Sciences, P.O. Box 80089, 3508 TB Utrecht ORCID ID: 0000-0003-1054-6683*

*[2]Email: o.a.savi@uva.nl Address: University of Amsterdam, Department of Psychology, Psychological Methods, P.O. Box 15906, 1001 NK, Amsterdam, The Netherlands ORCID ID: 0000-0002-9271-7476*

*[3]Email: A.D.Hofman@uva.nl Address: University of Amsterdam, Department of Psychology, Psychological Methods, P.O. Box 15906, 1001 NK Amsterdam, the Netherlands ORCID ID: 0000-0003-4269-5296*

*[4]Email: frecoomans@gmail.com Independent Researcher*

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

[5]Email: H.L.J.vanderMaas@uva.nl Address: University of Amsterdam, Department of Psychology, Psychological Methods, Postbus 15906, 1001 NK Amsterdam, the Netherlands
[6]Email: Gunter.Maris@act.org Address: ACTNext, 500 ACT Drive, Iowa City, IA 52245

## 1. Introduction

The societal expectations of educational data and learning analytics are high. As more and more educational institutes routinely use computerized tools for training and testing, enormous amounts of data on learning are collected. These data support the idea that "in the near future it will be possible to continuously assess and store the unfolding life history (trajectory in behaviour space) of each individual" (Molenaar, 2004, p. 216), and thereupon allow for a detailed study and targeted improvement of education. It should, for instance, be possible for teachers to create completely individualized educational programs based on the progress and learning difficulties of each student.

The role of learning analytics in shaping online learning systems is still emerging. In this paper, we contribute to this conversation by providing a case study of Math Garden (Klinkenberg, Straatemeier, & van der Maas, 2011), an online practice system for arithmetic. Math Garden aims to live up to the aforementioned promise by providing individualized computer-adaptive practice to over 400,000 primary school children in the Netherlands — by means of real-time ability estimates — and by giving teachers the tools to track the children's progress. In this case study, we first share design considerations, such as embedding learning analytics in the educational model, and then critically inspect the level to which those analytics are reliable. To this end, we determine the fit of the computer-adaptive model and explore sources of misfit. We finally share important considerations for the future of learning analytics.

We believe such a case study is particularly valuable, as designing learning analytics for such large-scale systems is not an easy task. Learning environments can have a sizable impact on education in general and individual students in particular. Moreover, interventions or design considerations in such systems, for instance based on learning analytics, may too have a significant impact on students' learning experiences, and must arguably be addressed with the same scrutiny as is demanded in traditional education.

To illustrate the reach, Math Garden involves almost 853,300,000 responses from over 452,000 K–12 children, distributed across 5,300 schools and many more household subscriptions, playing in 26 arithmetic domains totaling more than 37,000 different items. The rate at which items are answered is currently about 900,000 per school day. Then, to illustrate learning in Math Garden, Figure 1 shows the development of the domain *addition* over time (the Methods section explains the ability estimation procedure in detail). For each birth cohort (based on birth year), the development of average monthly performance is plotted, and nicely shows the development throughout a school year. After the school year, the development continues in the next class in the next year. Figure 1 illustrates how learning progresses through the years, and how classes compare to one another and over time. The graph includes almost 40 million responses from over six years of data. All analyses are performed using R (R Core Team, 2015).

Traditional psychometric methods, like classical test theory and item response models (e.g., Rasch, 1960; Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968), fall short in systems like these due to the scale and adaptive nature of these systems. Therefore, Math Garden utilizes a different approach, and analyzes student responses on the fly, while continuously updating the estimates of student abilities and item difficulties. In this case study, we draw upon the lessons we learned during a decade of analyzing Math Garden data.

Design considerations and the implemented computer-adaptive model are discussed in the Methods section. We scrutinize the fit of this model in the Results section by inspecting whether the predictions of the model are in accordance with the actual observed responses. Additionally, we determine the reliability of the model's estimates of item difficulties by comparing the estimates of parallel items (e.g., $n \times m$ and $m \times n$). We then take a deep dive into the many different possible causes of the small but significant amount of misfit observed in the system. We consider user-dependent responses processes, on both the global and local level: that is, the user-specific and item-specific strategies that children use for solving the items. The discussed sources of misfit primarily pertain to violations of the strict unidimensionality assumption that underlies the model.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

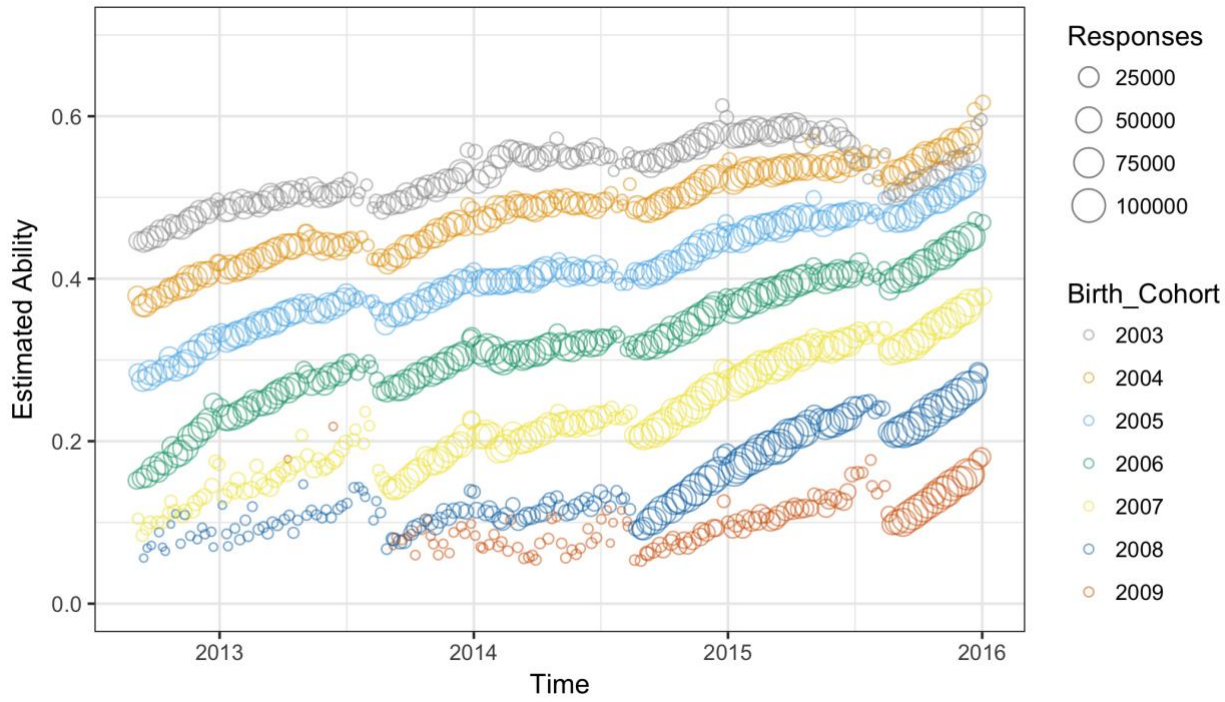


**Figure 1.** Growth in monthly average addition ability per grade, over a period of **6** years. Estimated ability represents the proportion of correct responses, if one responded to all addition items in the item bank. The number of responses in this graph totals $\mathbf{39,391,617}$, and the number of monthly responses can be seen to increase over time. For every new school year, the development of each grade is well visible. Also the continuation of progress over school years is clearly shown.

This multi-method approach exposes the fact that many facets play a role in educational systems that embed learning analytics into their educational model. In the Discussion section, we suggest that this embedded approach, combined with other active forms of learning analytics such as online experiments, might prove worthwhile as a first step towards a more coherent field. The central challenge in this approach is ensuring that the defined educational model works as desired — for which the following sections provide a case study.

## 2. Methods

### 2.1. Math Garden

Dating from 2007, Math Garden is a computer adaptive practice system for arithmetic items, mainly focused on K–12 (Klinkenberg et al., 2011; Straatemeier, 2014). Originally, it was designed to freely capture long and dense time series data for the microgenetic study of cognitive development in general, and mathematical development in particular. Due to popular demand, it was commercialized in 2009[1] and different domains were developed, such as the adaptive practice of languages (English and Dutch), statistics (Groeneveld, 2014; Klinkenberg, 2014), and typing (van den Bergh, Hofman, Schmittmann, & van der Maas, 2015). Each system hosts eight to 26 games that each train a distinct ability relevant to the domain.

Children who log in to Math Garden land on a personalized page with a garden and various plants (see Figure 2). Each plant represents a mathematical domain, such as addition, multiplication, or fractions. By clicking the plant, children start practicing that domain (see Figure 3 for an example). Plants grow and flourish when the corresponding domain is frequently practiced, while plants wither when a domain is neglected. In each practice session, a set of 15 items is sequentially presented for 20 seconds each. Depending on the domain, children either pick the correct response from a set of alternatives or respond in an open format. Children may, within certain limits, hit a question mark button to skip items that seem too difficult.

After each game, children return to the landing page where they can again choose a domain to practice. Math Garden adaptively matches children to items with an appropriate level of difficulty: an individual child who fails to solve an item or does so too slowly will receive easier items, whereas a child who succeeds within the expected time will receive more

[1] https://www.oefenweb.com

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

difficult items. Additionally, children can set the difficulty level (i.e., expected probability correct) themselves individually, to either easy (about 90% correct), medium (about 75% correct), or hard (about 60% correct). Adaptive item selection in the context of a practice system such as Math Garden is therefore quite different from item selection in computer adaptive tests (CATs). CATs optimize measurement efficiency by selecting maximally informative items for measuring ability (i.e., items with a probability correct of about 50% might be selected) to obtain maximum measurement precision within a limited set of items (e.g., van der Linden & Glas, 2002; Wainer, 2000). Adaptive practice systems on the other hand, choose items to facilitate learning and motivation, as discussed by Veldkamp, Matteucci, and Eggen (2011) and shown by Jansen et al. (2013). In Math Garden, no *optimal* item selection is currently attempted, but items are sampled, taking into account the preferred difficulty level of the learner and recent history of answered items to avoid recent items.


**Figure 2.** Landing page with a garden and various plants that represent mathematical domains. The smileys can be used to select the difficulty level. The buttons in the top right can be used to navigate to other parts of the environment, such as a bonus garden with more domains or a prize cabinet. Camera symbols communicate the availability of instruction videos.

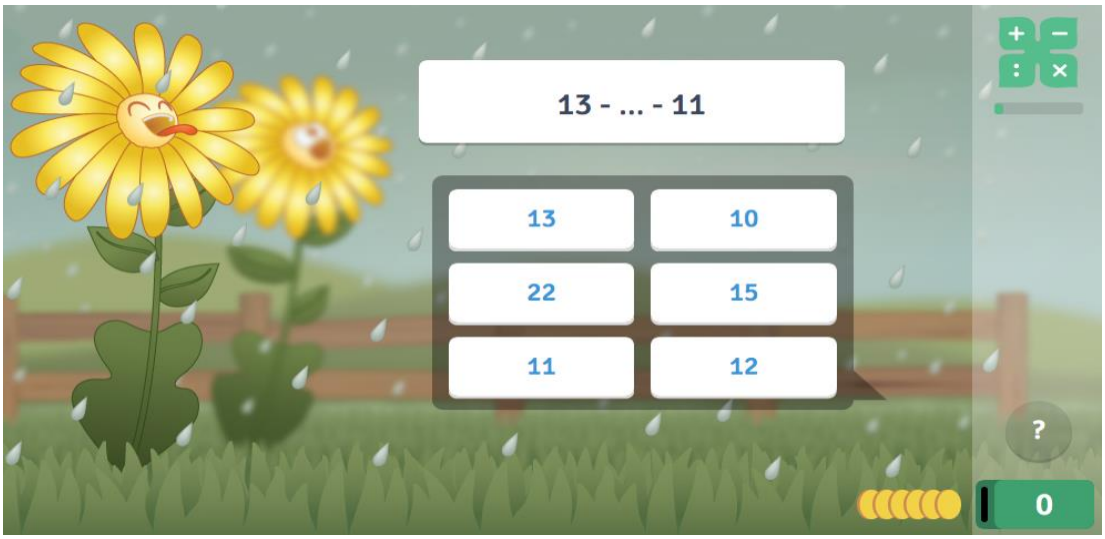


**Figure 3.** An item in the domain "series." Children must fill in the number that completes the incomplete series. The virtual coins indicate the remaining time. Children earn the remaining amount of coins if the answer is correct, and lose the remaining amount of coins if the answer is incorrect. The question mark can be used to skip the item.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

In order to anticipate the multidimensional structure of math practice, Math Garden is designed such that each of the games consists of a separate ability, hence is assumed to be unidimensional. This can be regarded as quite a conservative approach given the involvement of 26 games and thus as many dimensions. Many psychometric models assume unidimensionality, as does the Elo rating system (ERS). Hence, for each single game a separate rating scale is implemented. However, due to the large amount of data, it is still possible to distinguish different dimensions within a single game. This is demonstrated in the Model Fit Results section where misfit is discussed. Yet, the amount of bias introduced by this multidimensionality, together with all other sources of misfit, is believed to be limited, as is discussed in section 3.1.1. For each game within the Math Garden, two psychometric innovations are implemented: scoring rules and adaptive item selection, both discussed hereafter.

### 2.2. Scoring Rule

Scoring rules play an important role in such diverse domains as sports, games, educational testing, and recruitment. In all these domains, they are introduced to elicit specific behaviour that one somehow wants to quantify (e.g., answering correctly within a certain amount of time), thereby discouraging unwanted behaviour that can reduce the validity and reliability (also called the accuracy or dependability, cf. Cronbach, 1951) of the measuring procedure (e.g., guessing). See for example Lazer, Kennedy, King, and Vespignani (2014) for a general discussion on measurement in big data analysis, and Klinkenberg (2014) for an evaluation of the reliability and validity of the scoring rule. In large-scale computerized educational frameworks such as Math Garden, scoring rules additionally serve as a means to control the progression of the global system and to steer it towards a desired goal. To this end, it is important that the scoring rule is explicitly known and understood by the students, and that students act accordingly.

The scoring rule used in Math Garden, introduced by Maris and van der Maas (2012) and displayed in Figure 4, can easily be made explicit to the individual student. This scoring rule has the following form:

$$S_{pi} = (2x_{pi} - 1)(d - t_{pi}),$$

where $S_{pi}$ denotes the score earned by user $p$ after responding to item $i$, $d$ denotes the time limit and $x_{pi} \in \{0,1\}$ and $t_{pi} \in [0, d]$ denote, respectively, the accuracy and the response time of user $p$ on item $i$. In Math Garden the time limit $d$ is generally fixed at 20 seconds. The absolute value of the score $S_{pi}$ is determined by the remaining time until the time limit, $d - t_{pi}$, whereas the sign of $S_{pi}$ is determined by the accuracy $x_{pi}$.

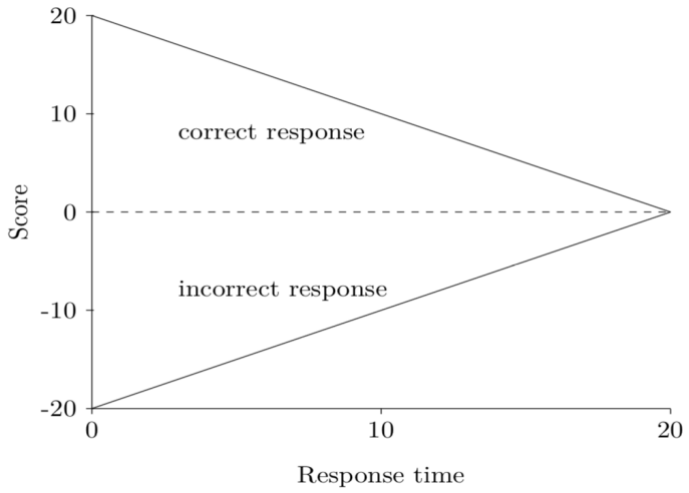


**Figure 4.** The Signed Residual Time scoring rule. If a user's response is correct, the score equals the remaining time until the time limit (shown by the top slope). If a user's response is incorrect, the score equals minus the remaining time until the time limit (shown by the bottom slope).

In this way, the scoring rule discourages fast guessing and imposes an explicit speed–accuracy trade-off (Wickelgren, 1977). The form of the scoring rule makes it easy to visualize the score to the individual user. At the start of an item the user sees a number of coins equal to the time limit in seconds, as visible in Figure 3. Each second one coin disappears. When a correct response is given the remaining number of coins is added to the total. In case of an incorrect response, it is subtracted. In Math Garden, children can collect these coins to buy virtual prizes. To allow users to omit an item that they do not know the

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

correct response to, without having to wait until the time limit has passed, a question mark button has been built in. By using this question mark button, a user can go to the next item directly, and earns a score of zero on the skipped question, though its use is now limited to constrain strategic behaviour in which students only try very easy items to maximize their points. Unless otherwise stated, all analyses presented in this paper are based on data from which these question mark responses are removed.

From the Signed Residual Time (SRT) scoring rule, Maris and van der Maas (2012) derived a response model. To estimate the response model's parameters on the incoming data streams from Math Garden, a rating system is implemented for each of the 18 games, facilitating real-time parameter updates, and driving the adaptive item selection discussed hereafter.

### 2.3. Adaptive Item Selection

To provide adaptive item selection, there is a need to determine what items are suitable to present to a specific student at a specific time. An algorithm based on the Elo Rating System (ERS) that both continually estimates the difficulty of the items and the ability of the students is used for this purpose.

The ERS has a history in the chess community, where dynamically changing abilities of chess players are expressed in Elo ratings (Batchelder & Bershad, 1979; Batchelder, Bershad, & Simpson, 1992; Elo, 1978). This provides a means to estimate dynamic ratings in setups that involve possibly massive paired comparisons. Hence, it is suitable for application in an educational context where item responses can be regarded as person–item paired comparisons, and we expect abilities and item difficulties to change over time (e.g., Klinkenberg et al., 2011; Brinkhuis, Bakker, & Maris, 2015; Pelánek, 2014; Wauters, Desmet, & Van den Noortgate, 2010).

To use the ERS in a computerized adaptive system like Math Garden, several modifications are required. First of all, the opposing player is replaced by an item $i$ such that a user $p$ responding to an item $i$ is considered a match between the user and the item. This match is won by the user if the response is correct, and won by the item if the response is incorrect. The ratings correspond to the user ability $\theta_p$ and the item difficulty $\delta_i$, the score corresponds to the SRT score (previous equation), which takes values in the interval $[-d, d]$, and the response model from which the expected score is computed is provided by the SRT model. After user $p$ responds to item $i$ and achieves SRT score $S_{pi}$, the user and item ratings are updated as follows (Klinkenberg et al., 2011):

$$\begin{aligned} \theta_p &\rightarrow \theta_p + K(S_{pi} - \mathcal{E}(S_{pi})), \\ \delta_i &\rightarrow \delta_i - K(S_{pi} - \mathcal{E}(S_{pi})), \end{aligned}$$

where $K$ is a scaling factor and the expected score $\mathcal{E}(S_{pi})$ is based on the current ability estimate $\theta_p$ and item difficulty estimate $\delta_i$:

$$\mathcal{E}(S_{pi}|\theta_p, \delta_i) = d\,\frac{\exp(2d(\theta_p - \delta_i)) + 1}{\exp(2d(\theta_p - \delta_i)) - 1} - \frac{1}{\theta_p - \delta_i}$$

where $d$ is the time limit. Brinkhuis and Maris (2009, p. 11) provide an intuitive visualization of how such updates work.

The ERS has two specific advantages that are beneficial in the context of adaptive practice. First, the method in which ratings are updated makes them self-correcting. In the equation above, the part $S_{pi} - \mathcal{E}(S_{pi})$ is simply the observed minus expected score. These differences facilitate the ERS to be self-correcting in its ratings — updates always steer in the right direction (e.g., a score that is higher than expected always gains points) — and the update size is related to the difference between observed and expected scores (e.g., for an unexpected correct response, the difference between observed and expected is quite high, and therefore the rating update is quite large, while for an expected correct response, the rating update is relatively small or can even be negative if the response given is too slow). This self-correcting feature makes the rating system quite robust: after every new response, ratings are updated in a sensible direction and hence adapt to changes in the underlying parameters. The $K$ factor in ERS functions is a scaling factor, and determines the size of the influence of the current response on the update of the ratings. A high $K$ factor allows for ratings to quickly adapt to changes in the underlying parameters, yet introduces noise, whereas a lower $K$ obtains smoother rating developments at the risk of adapting too slowly. This can be regarded as a classical bias–variance trade-off. Discussions on how to optimize $K$ can be found in Klinkenberg et al. (2011; Elo, 1978; Glickman, 1999, 2001), and Sonas (2005).

A second beneficial feature of this rating system is that it is iterative and computationally light. When the ERS was first introduced in the 1960s, updates could be calculated by hand, with the assistance of simple tables (Elo, 1978). The expected win probabilities depend on straightforward functions of estimated parameters, not on past data, and can be easily obtained.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Clearly, with the advent of big data, this allows for real-time calculations on possibly large streams of data, with little computational load. In the implementation of Math Garden, parameters are therefore updated in real-time as responses become available. Note that real-time updates of parameters with IRT models is challenging — see for example Veldkamp et al. (2011) for an approach on updating ability parameters, or Brinkhuis (2014, pp. 83–114) for a (time-intensive) approach to updating parameters on a daily basis.

The ERS allows us to obtain up-to-date estimates of both person ability estimates $\theta_p$ and item difficulty estimates $\delta_i$, which are continually adapted to possible changes. These ratings are used to facilitate many functions, such as adaptively selecting items at different difficulty levels, and providing teachers with child ratings and reference groups.

## 3. Model Fit Results

Having discussed the design considerations and embedded learning analytics in Math Garden's computer-adaptive system, in this section we evaluate its model fit. Relevant for both the field of learning analytics in general, and computer-adaptive practice environments in particular, we scrutinize model fit by exploring various causes of misfit. To this end, we use a variety of methods on very diverse sets of data from the Math Garden ecosystem.

### 3.1. Evaluation of Model Fit

We start by a general evaluation of the computer-adaptive Elo model that underlies Math Garden, specifically by evaluating the quality of fit of the model. Model fit is evaluated in two specific traditions. First, we use prediction accuracy from the field of machine learning. Second, we use reliability measures from the field of psychometrics.

#### 3.1.1. Prediction Accuracy

Figure 5 gives an indication of the quality of fit of Math Garden's computer adaptive practice model. It shows the amount of practice, and the extent to which it is able to predict a child's responses. To be more precise, the figure shows for one particular child over the course of three years the difference between the observed and expected SRT scores, normalized over $d$, to every single item (s)he attempted in the fields of addition, subtraction, and multiplication. These differences can be interpreted as a proxy for model fit for this person. Hence, the closeness of these differences to the zero-lines correspond to good prediction, and therefore good model fit. The RMSE decreases from .45 ($n = 5{,}804$) to .29 ($n = 4{,}312$) for addition, and from .54 ($n = 4{,}369$) to .31 ($n = 3{,}305$) for subtraction, comparing 2012 and the rest. The total RMSE is .41 ($n = 20{,}392$).

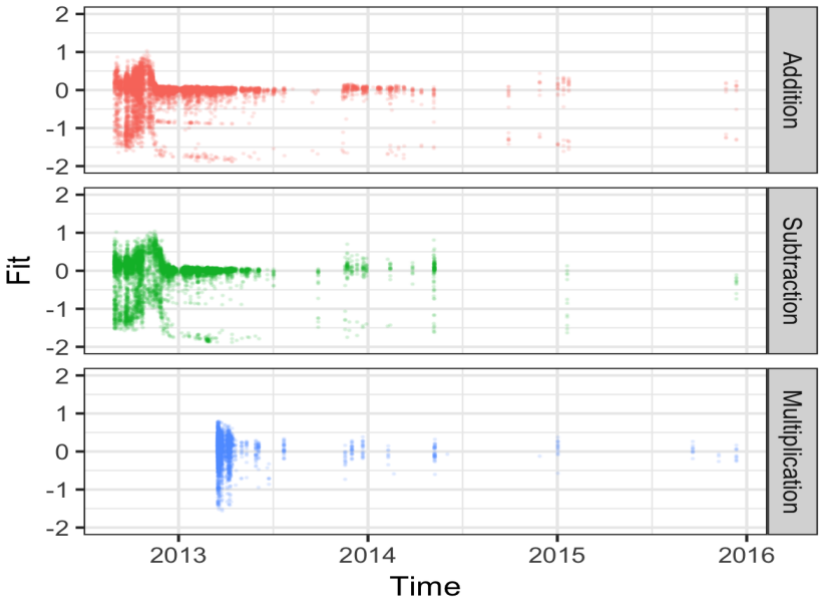


**Figure 5.** The development of model fit for one particular individual over time, for the domains addition, subtraction, and multiplication. For every single response over the course of three years ($\boldsymbol{n = 20{,}392}$), the differences between the observed and expected SRT scores are shown, normalized over the time limit. The smaller this difference, the more accurate the expected score. For this individual, fit can be seen to improve over time on all three domains. The onset of practice differs between domains, and the amount of practice drops for these domains after May 2014. Some bimodality can be observed (partly due to guessing).

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

In addition, one can see intensive practice starting in September 2012, and lasting through June 2013. Observations span a couple of years, since in Math Garden users are encouraged to revisit domains occasionally. Practice levels of this child and for these domains decline sharply after May 2014. First, narrowing down to the quality of fit, we see that after some initial phase, the difference between observed and expected responses tends to get centred closely around zero. At the onset of a new domain there is quite some noise, which reduces after some time. Since the estimated item difficulty parameters are readily available, the improvement in fit is not only due to better parameter estimation. This user increasingly conforms to the scoring rule and the response model that goes with it, which can also be observed by the relatively fast increase in fit in the multiplication domain. Hence, the estimated model parameters facilitate a good prediction. That Elo ratings can provide good prediction accuracy is not unique to Math Garden, and for instance also shown by Nižnan, Pelánek, and Řihák (2015).

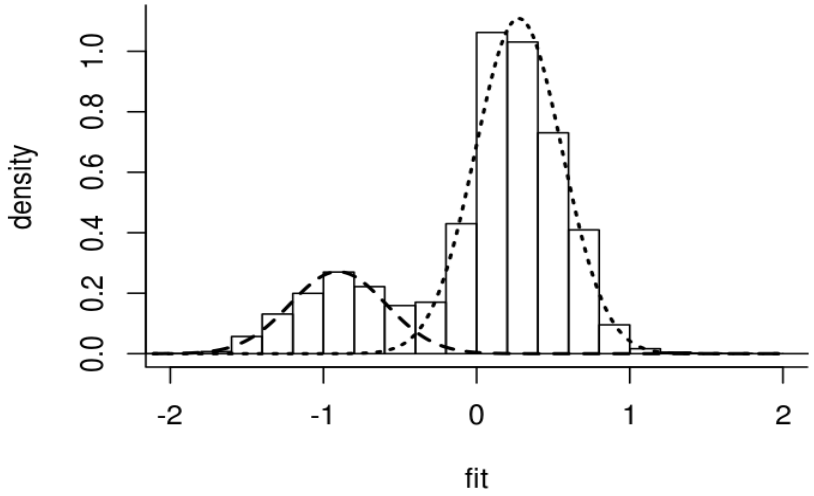


**Figure 6.** Histogram of the difference between the observed and expected SRT scores, normalized over the time limit, for all $\mathbf{463,729}$ Math Garden responses on May 26, 2015, excluding skips. Overlaid is a fitted mixture of two normal distributions. The smaller distribution on the left contains $\mathbf{21}\%$ of the observations at $\mathcal{N}(\mathbf{-0.90, 0.31})$ (dashed), and the larger distribution on the right contains $\mathbf{79}\%$ of the observations at $\mathcal{N}(\mathbf{0.28, 0.28})$ (dotted). The mean difference between observed and expected scores is close to zero ($\mathbf{0.02}$). The smaller distribution on the left corresponds to person-item interactions, where the expected result was correct, but the observed score was fast and incorrect — typical for typing errors and guessing.

Figure 6 provides another representation of the difference between observed and expected responses, this time for a large group of students. For one particular day, May 26, 2015, we have selected all responses for all games in Math Garden. As this day is situated at the end of the school year, we expect few new students and hence expect fit to have converged for this group, e.g., see Figure 5 for an improved fit near the end of the school year. On this day, 13,608 students provided 463,729 responses to 10,983 items on 17 different games. The differences between the observed and expected scores, normalized over time limit $d$, for all these responses are provided in Figure 6. The mean difference is close to zero ($\sim 0.02$), which means that the ERS appears to do a good job at adjusting the expected scores toward the observed scores.

As the expected responses are best guesses at the time the actual responses are observed, the achieved accuracy indicates a significant amount of control on the dynamics in the environment. Nonetheless, Figure 5 also shows a constant stream of responses for which the observed score is not close to the expected one, and in Figure 6 one can clearly see that the histogram is bimodal.

We estimated a mixture of two normal distributions on this data, using the mixtools R package (Benaglia, Chauveau, Hunter, & Young, 2009)., and obtained a smaller distribution with 21% of the observations at $\mathcal{N}(-0.90,0.31)$, and a larger proportion of 79% at $\mathcal{N}(0.28,0.28)$. The smaller distribution seems to collect all sorts of unexpected errors, such as typing errors, or fast guessing (Wang & Xu, 2015). When rather easy items are selected for the students, which have positive expected value, a quick error results in a quite large negative observed score (Figure 4), resulting in a (large) negative difference between observed and expected scores. These errors can also be observed in Figure 5, where for all three domains points can be seen hovering at the lower end of the panels. Since these errors are asymmetric, they introduce some bias in the estimated expected scores. Such bias might be removed, for example by disregarding quick incorrect responses in updates of

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

the ratings. The larger component $\mathcal{N}(0.28,0.28)$ includes responses that conform to the SRT score model (i.e., excluding guessing, typing errors, etc.), and allows us to estimate the RMSE of prediction to be 0.28, across all Math Garden games, excluding quick incorrect answers. Further considerations of fast and slow processes are discussed in section 3.2.

#### 3.1.2. Reliability of Parallel Item Development

In addition to prediction accuracy, the quality of fit of the adaptive system can be investigated by comparing the temporal development of the ratings of parallel items. Such temporal developments can be interpreted as a measure of reliability of the measurement, since similar items should have a similar (development of) difficulty. Parallel items look different superficially but share a number of features, which make them more or less equivalent. See Brinkhuis et al. (2015) for an approach to detect differential development of items of pairs. An example of such parallel items is provided by mirror items like $2 \times 9$ and $9 \times 2$. In the left panel of Figure 7, the temporal development over a period of 3.5 years of the daily average ratings of two pairs of parallel items from the multiplication domain in Math Garden are displayed. It is clear that for both pairs the ratings of the parallel items remain very similar over time: their temporal rating evolutions overlap to a large extent, though the ERS allows for estimating their item difficulties independently.

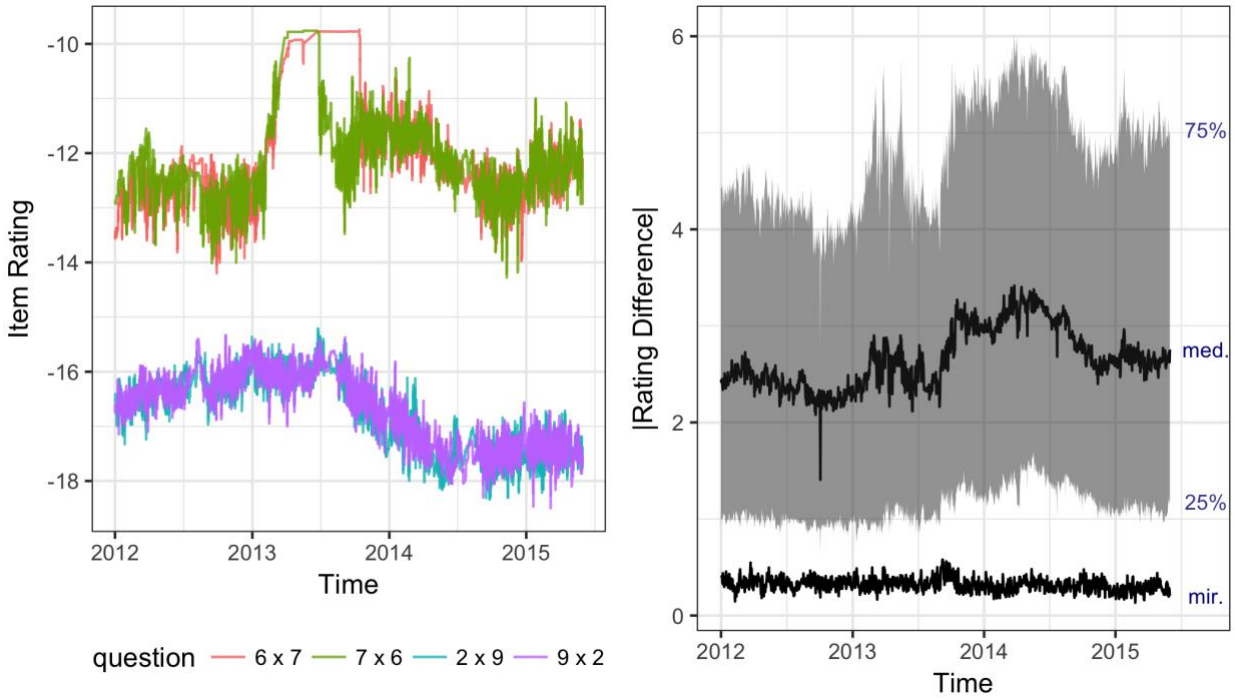


**Figure 7.** (Left) Temporal development of the daily average rating of 2 pairs of parallel items from the multiplication domain. The lower pair of ratings constitutes the items $\mathbf{2 \times 9}$ and $\mathbf{9 \times 2}$ and the upper pair constitutes the items $\mathbf{6 \times 7}$ and $\mathbf{7 \times 6}$. Item ratings are non-transformed Elo ratings. (Right) Temporal development (grey area) of the distribution of the absolute item pair-rating difference of all 4,005 item pairs by the 90 non-symmetric items in the multiplication table. The absolute item pair-rating difference of the 45 mirror item pairs can be found at the bottom of the figure. Med. refers to the median of the distribution of non-mirror items (with 25%–75% boundaries) and mir. refers to the median of the mirror items.

The right panel of Figure 7 generalizes these findings to all 45 parallel item pairs that can be formed by the 90 non-symmetric items in the multiplication table. For every day in the 3.5-year period, the median, 25%-, and 75%-quantiles are determined of the absolute value of the item pair differences in daily average rating for all 45 item pairs. To put this in the right perspective, the figure also displays the temporal development of the median of the absolute item pair-rating differences computed over all 4,005 item pairs that can be formed by the 90 non-symmetric items in the multiplication table. This figure makes it quite clear that the ratings of parallel items remain much closer over time than the ratings between two generic items. Even though other single-digit multiplication items can be equally difficult, the consistent small differences between mirror items is an indicator of the reliability of these ratings. See van der Ven, Straatemeier, Jansen, Klinkenberg, and van der Maas (2015) for considerations on the difficulties of single-digit multiplication, and when items cannot be considered mirror items.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

Taking the above results together, both the prediction accuracy and the item ratings of parallel items suggest that the computer-adaptive architecture is able to create a considerable amount of stability within such a complex dynamical system. Given this achieved stability, the observed data collected with the system allows for a detailed look at the cognitive processes used in learning arithmetic. However, we also observed a certain amount of misfit, which we investigate next.

### 3.2. Diagnosis of Model Misfit

The methods in the previous section give different perspectives on the quality of fit, and give rise to further — more specific — explorations aimed at diagnosing misfit. Figures 5 and 6 display a number of promising results that indicate a good working of the mechanics underlying Math Garden. However, the dashed mixture component in Figure 6 clearly indicates a set of responses for which there are substantial differences between the observed and expected scores. These types of responses indicate alternative behaviour (e.g., typing errors or fast guessing), which may lead to an incorrect assessment of the ratings, and to misfit of the response model.

A good working of the system requires the detection of the sources of this misfit after which appropriate steps can be taken to properly deal with these. First, we diagnose misfit by investigating different response processes (section 3.2.1). Second, we analyze local item strategies (section 3.2.2). Finally, we provide a short overview of other sources of misfit, illustrating the complexity of identifying and correcting sources of misfit (section 3.2.2).

#### 3.2.1. Global Response Processes

An attempt to assess the misfit of the SRT model and to situate the model in a more generalized framework of speed–accuracy response models can be found in Coomans, Hofman, Brinkhuis, van der Maas, and Maris (2016). In that paper the quality of fit of several of these models, including the SRT model, is investigated in the simplest possible non-trivial setup: persons try to solve two problems only; it is registered whether or not their response is correct and whether their response time is faster than half the time limit or slower than half the time limit. Hence, in this setup there are four different ways a person can answer a single item (fast and correct, slow and correct, slow and incorrect, and fast and incorrect) and 16 different ways a person can answer a pair of items. This simplistic setup is advantageous because:

- It gives access to data from a large number of item pairs, spanning such diverse subject areas as basic arithmetic, language learning, and intelligence-related problems, with large numbers of independent observations per item pair.
- Different speed–accuracy response models predict qualitatively different probability distributions of the 16 possible response patterns in a population of test takers. By inferring these distributions empirically by using, for example, Math Garden data, we can easily get a handle on the allowed speed–accuracy trade-off mechanisms.

To give an example of the analysis done in Coomans et al. (2016), reconsider the item pair $9 \times 2$ and $2 \times 9$, previously discussed in Figure 7. We obtained the response patterns of 13,152 persons who responded to this pair of items within one day, and collapsed the response times in two categories: response times smaller than half the time limit are classified as slow, others as fast. The resulting data is summarized in the contingency table displayed as Table 1.

**Table 1.** Item pair contingency table for the items $\mathbf{9 \times 2}$ and $\mathbf{2 \times 9}$, constructed from $\mathbf{13,152}$ persons who responded to the item pair within a single day (over the period 2011-03-01 to 2015-06-29). The cells are numbered using superscript. The cells 1, 4, 13, and 16 constitute the events for which both responses on the item pair are fast. The cells 6, 7, 10, and 11 constitute the events for which both responses on the item pair are slow. All remaining cells constitute the events for which the speed of both responses on the item pair differs.

| | | $\mathbf{2 \times 9}$ | | | |
|---|---|---|---|---|---|
| | | incorr./fast | incorr./slow | corr./slow | corr./fast |
| $\mathbf{9 \times 2}$ | incorr./fast | $313^{1}$ | $107^{2}$ | $137^{3}$ | $434^{4}$ |
| | incorr./slow | $124^{5}$ | $98^{6}$ | $153^{7}$ | $230^{8}$ |
| | corr./slow | $132^{9}$ | $130^{10}$ | $684^{11}$ | $1221^{12}$ |
| | corr./fast | $440^{13}$ | $190^{14}$ | $1211^{15}$ | $7550^{16}$ |

In Coomans et al. (2016) it is demonstrated that the SRT model constrains the expected frequencies of the response patterns on anti-diagonals (9,6,3), (13,10,7,4), and (14,11,8) to be monotonically increasing or decreasing along these anti-diagonals. However, Table 1 is clearly incompatible with these predictions: the frequencies of the events on anti-diagonals (9,6,3), (13,10,7,4), and (14,11,8) are not monotonically increasing or decreasing along these anti-diagonals, but instead exhibit a dip (along (9,6,3)), a dip (along (13,10,7,4)), and a peak (along (14,11,8)). The same features are found for

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

numerous other item pairs in different domains, for all of which there are a great many observations that can be easily extracted from the Math Garden database.

Coomans et al. (2016) concludes that these features cannot be accounted for by simple "one-process" models, such as the SRT model, and that a more complex model is needed. Therefore, they consider a "two-process" model developed in Partchev and De Boeck (2012) and which explicitly distinguishes between fast and slow responses, showing that this model results in a better fit than the more parsimonious SRT model, which does not make such a distinction. A similar conclusion was reached in Hofman, Visser, Jansen, Marsman, and van der Maas (2017).

#### 3.2.2. Local Response Strategies

We will now turn to an example where the adaptive nature of Math Garden steers towards undesired behaviour, ultimately resulting in misfit of the model. This example was encountered in the balance-scale task (Inhelder & Piaget, 1958), implemented in Math Garden in 2010. In this task, children predict the movement of a balance-scale (see Figure 8), with a varying number of blocks on each peg and varying distance between the blocks and the fulcrum. The task is famous for the interesting (erroneous) strategies used by children (and adults). To discriminate between these strategies, Siegler (1976) classified items to different item types. Simple items are included to discriminate between children who use a simple strategy based on only one dimension; counting only the number of weights or only looking at the distance between the location of the blocks and the centre of the scale. Next to the simple items, complex items are added where children need to integrate both the weight and the distance information to correctly solve the item.

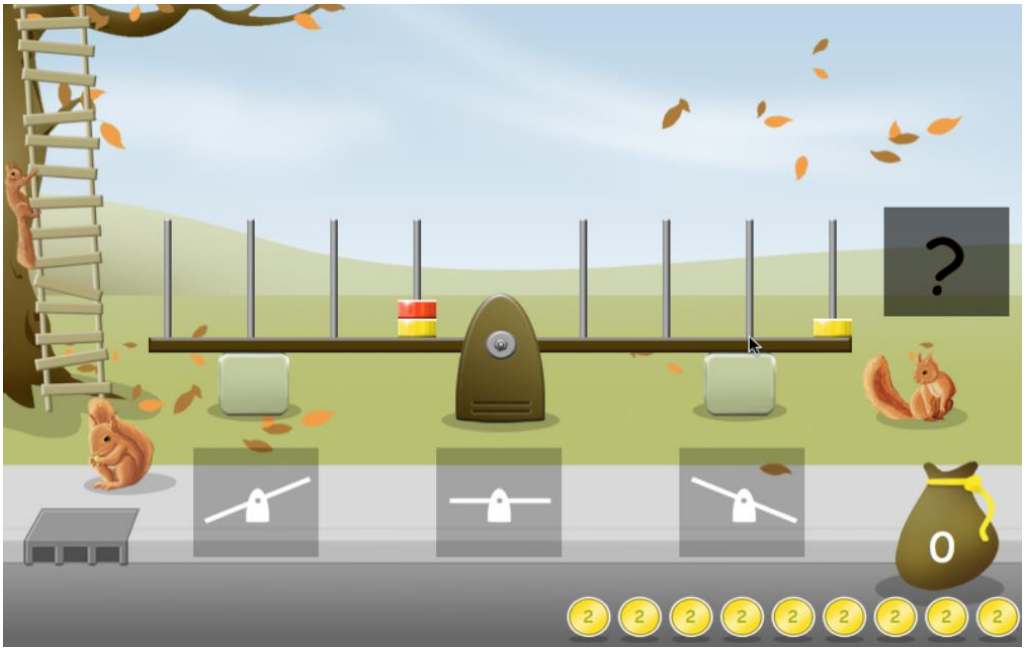


**Figure 8**. Example of the balance-scale task, as implemented in Math Garden. The coins reflect the used scoring-rule: if a correct/incorrect response is provided, the coins are added/subtracted to/from the child's virtual savings. Children respond by clicking the left, middle, or right picture of the balance scale — depicting the three possible states of the balance scale — or by clicking the question mark button.

In the first implementation of the task, the adaptive item selection was based on the differences between estimated item difficulties and user abilities, as in all other Math Garden domains. Interestingly, when items are selected based on the Elo parameters, the collected estimated ability ratings for individual users show jumps between qualitatively different strategies. However, when items are selected in a fixed order, no such development is present (see left-panel of Figure 9 for both patterns). Also, a closer look at the responses to items (see right-panel of Figure 9) reveals that item responses are clustered when items are selected based on Elo ratings (groups of correct and incorrect responses are visible).

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

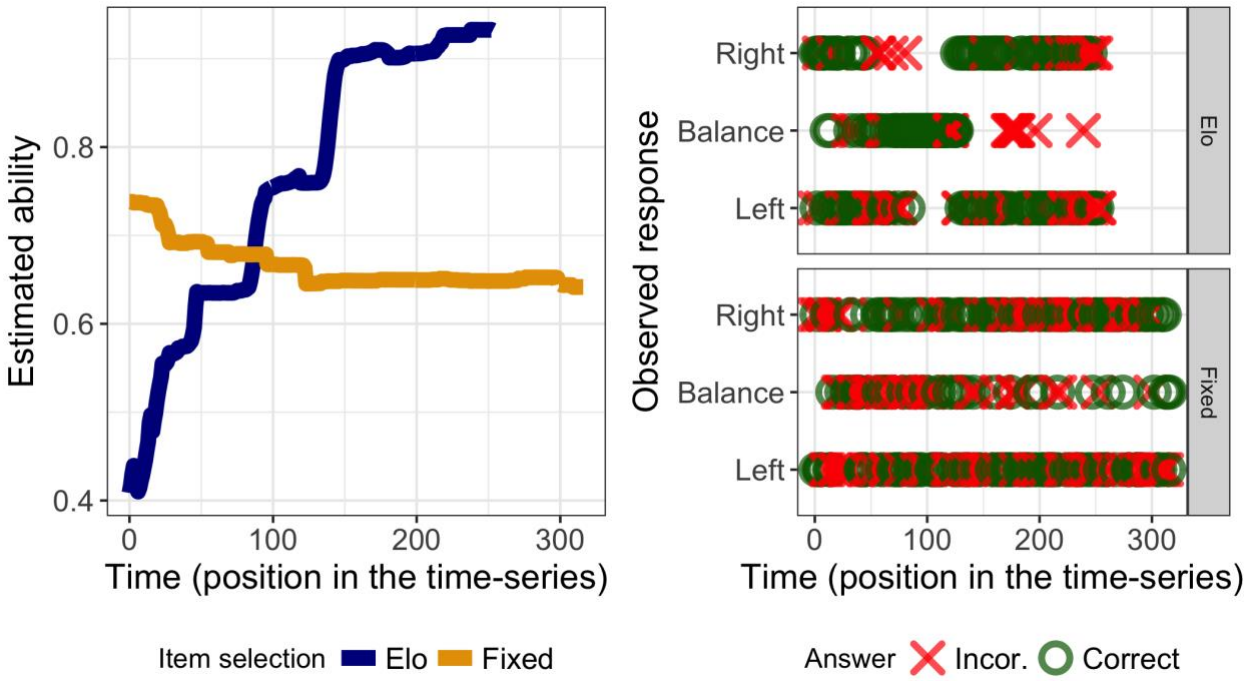


**Figure 9.** The development of responses to the balance-scale task, for a single child. The left-panel shows the rating development of a single user. When items are selected with the adaptive item selection based on the Elo estimates, large jumps in rating are observed, but when items are selected with a fixed sequence, a decrease in rating is observed for all users. The right-panel shows the responses (left, balance, or right), and whether the response was correct or incorrect, for both the data collected with the adaptive item selection procedure (clustering is visible) and fixed item selection procedure (no clustering visible).

This clustering has an intriguing cause. The users seem to develop a local strategy that only works on the cluster of items presented at that specific moment. For example, a user might recognize that the response "balance" is not correct for the first few items and learns that the balance response is always incorrect. Between item position 25 and 50 some of these items are presented but made incorrect, hence the ability estimate does not increase in the left panel of Figure 9. Since the system adapts the item difficulty estimates based on these responses, the difficulty estimates of these items increase and of the remaining items decrease. This results in an automatic clustering of items for which this local strategy fails versus items on which the strategy succeeds. After some incorrect responses, and receiving feedback, this child learns that he/she should provide only balance responses, as can be seen around item 80. This results in an increase in the ability estimate, what eventually results in the selection of items of yet another type (around 130), and a new local strategy seems to be learned.

In this example, the dynamic estimation of the item and user ratings, in combination with local strategies, result in dynamics that reinforce the reward of developing erroneous local strategies. Importantly, in this situation an ability estimate is based on the local cluster of items that the user has practiced, and does not generalize to the other clusters. This violates the assumption of unidimensionality and results in misfit.

To solve this undesired state of the system we intervened on the item selection by presenting a fixed order of items to all children, thus making the system less adaptive. The collected data and estimated ratings in the new implementation of the task showed large deviations compared to the first implementation. For example, the observed responses and ability ratings do not show a clear developmental pattern when items of different types are mixed (see the lower-panel of Figure 9). Although changing the item selection resolved the development of local strategies, still large discrepancies are found in the strategies used by children in Math Garden compared to strategies invoked by more traditional paper-and-pencil tests (Hofman, Visser, Jansen, & van der Maas, 2015). Clearly, such interactions between the content of the domain and the adaptive algorithm are not easily foreseen and require careful investigation into sources of misfit.

#### 3.2.3. Other Sources

The previous example illustrates that complex learning systems can have undesirable side effects, and that one should be on guard for unexpected behaviour in different forms. Having diagnosed sources of misfit in alternative response processes and

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

specific item strategies, in this section we briefly identify four more sources of misfit that are currently active areas of research in Math Garden. This collection of sources further demonstrates the rich variation in sources of misfit, the diverse set of methods required for their diagnosis, and hence the complexity in reducing misfit.

One source of misfit can be found in the (lack of) adherence to the scoring rule. A good working of the system implies that its users respond in accordance with the scoring rule, i.e., that their ability is reflected in the score that they earn. This is ensured in part by the form of the SRT scoring rule which strongly discourages guessing and thus prevents low ability users to earn scores that are too high and do not correspond to their actual ability. However, despite this explicit penalizing of fast incorrect responses, a substantial amount of guessing remains. Moreover, the particular form of the scoring rule can have a negative effect on less confident, yet high ability users. They might be scared by the high stakes associated to fast responding and produce a slow response resulting in a score that is too low for their actual ability. For these reasons it is important to develop methodology that enables an evaluation of scoring rules to find out if the (majority of) users conform to the SRT scoring rule, as discussed by Klinkenberg (2014).

A second source of misfit is very much related to the issue of users not adhering to the scoring rule. As mentioned in the section, children earn virtual coins by giving correct responses, and faster responses yield more coins. Some children who aim to maximize the number of collected coins are observed to quickly skip problems that they deem too difficult to answer within a short time. They quickly use the question mark button to proceed to the next problem, as they're not penalized for doing so, and wait for an item that they can quickly answer correctly. This way, they somewhat circumvent the adaptive item selection by only choosing items that yield the most coins. Ultimately, this strategic behaviour results in subtle misfit, as these children's abilities cannot be assessed correctly. For assessing such misfit, standard errors of estimates in the ERS would be beneficial, as explored by Brinkhuis and Maris (2010).

Interestingly, a solution to this issue was implemented in the Math Garden ecosystem. Savi, Ruijs, Maris, and van der Maas (2018) explain how a large online randomized experiment revealed that a simple delay in making the question mark button available, decreased the number of question marks used, and increased the amount of effort put into the children's responses. The development team of Math Garden has subsequently implemented such a delay throughout their ecosystem. The degree to which this intervention helped decrease misfit in the adaptive system is a subject of study.

For a third source of misfit, we investigate single-person-by-item time series. The size of the Math Garden data allows investigation of development in a new level of detail; that is, the individual development of accuracy, including the error types and response times, on a single item (Klinkenberg et al., 2011). These time series show interesting patterns from a developmental perspective and allow testing, for example, of whether learning of one set of items is related to learning another set of items. To illustrate the different patterns, we selected three of these series. Figure 10 shows the development of three different users on three different items ($5 + 1$, $3 + 4$, and $4 \times 3$) over a long period, with a maximum of 136 weeks.

The upper panel shows the responses of a child who learns to add five and one (and the parallel item), in three different stages. In the first stage, until position 25, he or she provides incorrect responses, mostly answering five. Thereafter, in the second stage, the correct solution is learned. In the third stage, from position 38 onward, the observed response time decreases indicating a more efficient strategy or faster sampling from memory (Ashcraft, 1982) compared to the previous stages.

On the other hand, the middle panel shows a time series of a child who does not learn $3 + 4$, while practicing this item for 61 times. Some correct responses are stated, but these are alternated with errors. These errors provide insight into the highly variable cognitive process of this child over time. The child alternates responses that can be labelled as close misses (6, 8 and 9) and responses labelled by wrong multiplication operand strategy (12). This highlights possibilities for tailoring instruction and feedback to misconceptions of a child as detailed as to a single item.

The development depicted in the lower panel shows a different pattern. This child starts with fast question-mark responses. Around position 22, he provides (too) slow responses, and seems to learn during this period the correct response. In the last part of the series he gives correct responses to this item. The dynamics of this child also highlight the differentiation between fast and slow processes, discussed previously.

The quantification of these developmental patterns, and the connection between multiple single-item time series, is ongoing research. Especially the connection between different time series provides insight into an important type of misfit. That is, the possible presence of item clusters within a certain domain, see for example Pelánek, Papoušek, Řihák, Stanislav, and Nižnan (2017) or the previous discussion of the balance-scale task. These possible item clusters show that learning a subset of items is strongly related to some items, whereas it is unrelated to other items within the same domain. These clusters provide insights in qualitative differences between the solutions strategies and learning patterns of children. Furthermore, the inclusion of these possible clusters in the measurement model can reduce the amount of misfit.

The fourth and final source of misfit is captured in the just mentioned error responses. Error analyses provide an interesting direction for investigating misfit, because information about individual cognitive processes is contained in the

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

types of errors that students make. More specifically, since different students may have different misconceptions, and different items are susceptible to different misconceptions, error analyses can help detect violations of unidimensionality.

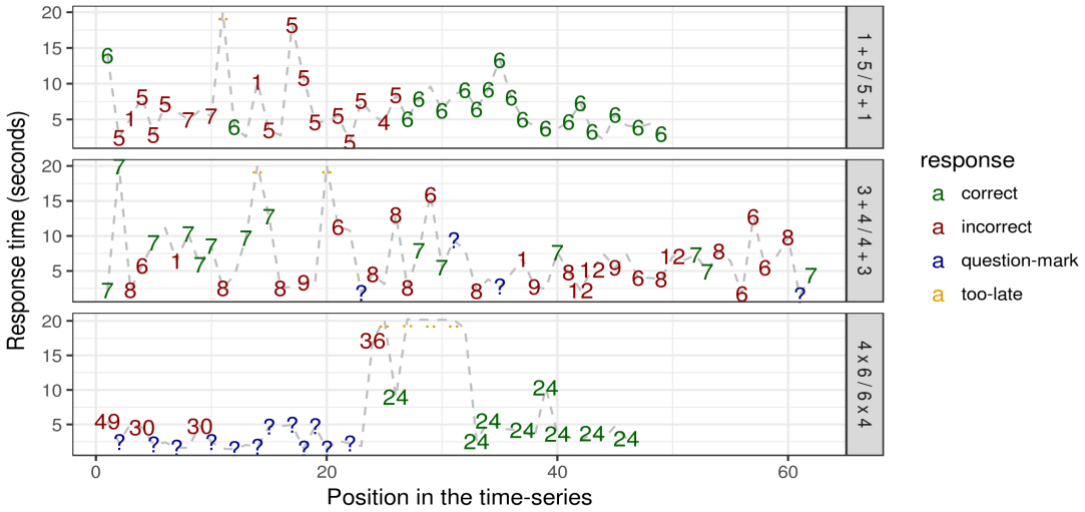


**Figure 10**. The development of both accuracy and response time, for responses to three different items (and their parallel items). For each item, the responses by one particular individual over time are shown. The ordered position in the time series are on the horizontal axis, which can span a maximum 62 responses. The upper and lower panel show a three-stage pattern, moving from mainly incorrect responses, to slow correct responses, to fast correct responses.

An example of error analyses is shown in Figure 11. It shows the frequencies of the most made errors of the same set of items as depicted in Figure 7. Different aspects are highlighted by this plot. First, the observed correspondence in the error frequencies between the two sets of parallel items supports the reliability of the system. Second, the most frequent error for each parallel item pair seems to indicate different processes. The response 19 to the item $2 \times 9$ implies a mistake in counting, since children missed the correct response by one. Whereas the response 56 to the item $6 \times 7$ implies an operand relevant mistake, since the answer is correct for another multiplication problem (Straatemeier, 2014, pp. 99–128).

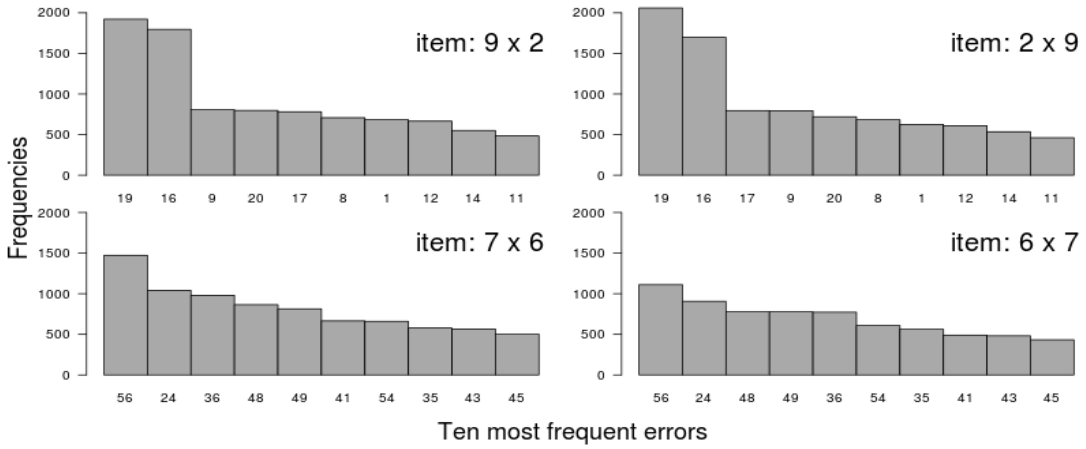


**Figure 11.** Frequencies of the ten most frequent errors on the items **2 × 9** and **6 × 7**, and their parallel items.

However, the classification of observed errors to error types is often ambiguous (Brown & VanLehn, 1980), and is therefore an active area of research in Math Garden. Take for example the incorrect response 18 to the item $9 \times 9$. This error can indicate that this child (1) adds instead of multiplies (wrong operand), (2) incorrectly reverses 81 to 18, or (3) states the response to an item from within the same table (operand related error). To solve this issue in error-classification, Straatemeier (2014, pp. 99–128) introduced and compared multiple classification methods (two literature-based approaches and four data-based approaches) that can be used with the availability of big data to uncover what types of errors a child makes on a particular item. Using the weighted frequency rule, more than 80% of the 1,104,865 errors can be classified as

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

coming from a certain strategy, and distinctions in age can be made. Importantly, since these errors provide information about which erroneous strategies children use to provide a response, the classification of these errors can provide a valuable tool in educational computer programs, as it allows for providing personalized feedback.

## 4. Discussion

The Math Garden ecosystem, like other large-scale learning environments, contributes to what Molenaar (2004) described as an opportunity to "continuously assess and store the unfolding life history (trajectory in behaviour space) of each individual." However, although the phrase nicely catches the opportunities of today's educational data, it fails to draft the desired way forward. In this final section, we reflect on the research discussed in the current paper, and give important considerations for the future of learning analytics.

### 4.1. Active analytics

A primary characteristic of learning systems is their educational objective, and this objective should have a central place in learning analytics. Savi, van der Maas, and Maris (2015) argue that reaching a desired educational objective requires one to first accurately track a student's development, and subsequently map each student's learning route. That is, the spectrum of each student's ability should be assessed and tracked over time, such that an accurate learner model is created. This learner model may for instance encompass the discussed ability measures for various scholastic domains, or possibly some diagnosed misconceptions, and should give rise to the creation of an optimal learning path for this particular student.

To this end, we believe that learning analytics should be an *active* exercise. Rather than passively collecting analytics about a learning environment, learning analytics must actively help direct a student towards his or her educational objective — such as effortful practice on the level of the individual child in the case of Math Garden. Math Garden applies active forms of learning analytics on multiple levels. First and foremost, as laid out in the section, it utilizes *embedded learning analytics*: the Elo rating system at the core of Math Garden estimates item difficulties and user abilities on the fly, and dynamically steers each student's learning experience in the desired direction.

Moreover, as we show in the section, making sure a learning system optimally directs the student towards the intended goal additionally requires active development. In this paper, we took model fit as the primary approach, and showed how different data selections and different methods shed diverse lights on the problem of misfit. We showed that although in general the adaptive system both accurately predicts student responses and reliably estimates item difficulties, these analytics may be biased by systematically misfitting responses. Multiple sources of this misfit were discussed, such as the possibly distinct processes underlying observed responses and local response strategies for subsets of items. Finally, the rich diversity of possible sources became evident when we discussed four more explorations of misfit, including a diversity in possible error patterns, and unexpected and undesired consequences of the used scoring rule.

The nevertheless good fit of the system illustrates that embedded learning analytics can help track and direct the development of an individual student. We hope to have conveyed that model fit can be seen as a central endeavor in learning analytics, with implications for very diverse parts of a learning system. Moreover, active analytics, such as the embedded learning analytics employed in Math Garden, need to assure that the system and its users reach their educational objectives.

Besides the embedded learning analytics, we believe a second form of active analytics deserves careful consideration: *experimentation*. The different sources of misfit in the section illustrate that without careful supervision, the ecosystem may move towards an unintended or even undesirable goal. Moreover, the necessary continuous maintenance of a large-scale online learning system like Math Garden unmistakably changes the system in both intentional and unintentional ways. In such a goal-directed system, these changes can alter the degree to which the goal is reached. Experiments serve to detect how an intervention alters the complex system, and to make sure it does not behave in unintentional and possibly detrimental ways.

An experimental method particularly suited to large-scale online learning systems is the online randomized controlled experiment, commonly known as the A/B test (Savi, Williams, Maris, & van der Maas, 2017). In the section, we briefly discussed one such experiment, aimed at preventing undesirable strategic responses that increase the misfit of the adaptive system. Additionally, besides using experiments to evaluate the mechanics of a learning system, experimental comparisons of pedagogical interventions can provide additional leverage. The learning sciences provide a wealth of possible interventions targeted at achieving learning gains, and often well suited for testing. Similarly, large online educational systems provide an exceptional testing ground for such interventions.

## 5. Conclusions

Although a vast share of research on learning is conducted within the safe boundaries of confined experiments, that is not

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

where the actual everyday learning happens. Everyday learning happens in vivo — in a complex, dynamic, ecological system. Such a system is inherently difficult to track, let alone deliberately navigate towards a desired goal. Fortunately, an ever-increasing worldwide accessibility to the internet and serious efforts to scale learning technologies, increasingly succeed to unlock the big data of learning. These data, with an unprecedented granularity, combined with advanced methods, are now starting to provide a window into the complexity and dynamics of learning in vivo. In this paper, we reported on a decade of experience from one such system, Math Garden. We described what we have learned and how we are still learning from a system that develops while we observe learning as it happens.

## Acknowledgement

Matthieu Brinkhuis and Alexander Savi contributed equally to this work. Frederik Coomans and Abe Hofman contributed equally to this work.

## Declaration of Conflicting Interest

Han van der Maas is full professor of Psychological Methods at the University of Amsterdam and founder of Oefenweb, the company that operates Math Garden. Abe Hofman is postdoctoral researcher at the University of Amsterdam and part time employed as a researcher at Oefenweb. However, this study does not aim to demonstrate the effectiveness of training in Math Garden.

## Funding

The publication of this article received financial support from the Netherlands Organisation for Scientific Research (NWO), grant numbers 314-99-107 and 406-11-163.

## References

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review*, *2*(3), 213–236. http://dx.doi.org/10.1016/0273-2297(82)90012-0

Batchelder, W. H., & Bershad, N. J. (1979). The statistical analysis of a Thurstonian model for rating chess players. *Journal of Mathematical Psychology*, *19*(1), 39–60. http://dx.doi.org/10.1016/0022-2496(79)90004-X

Batchelder, W. H., Bershad, N. J., & Simpson, R. S. (1992). Dynamic paired-comparison scaling. *Journal of Mathematical Psychology*, *36*, 185–212. http://dx.doi.org/10.1016/0022-2496(92)90036-7

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, *32*(6), 1–29. http://dx.doi.org/10.18637/jss.v032.i06

Brinkhuis, M. J. S. (2014, December). *Tracking educational progress* (PhD thesis). University of Amsterdam. http://hdl.handle.net/11245/1.433219

Brinkhuis, M. J. S., Bakker, M., & Maris, G. (2015). Filtering data for detecting differential development. *Journal of Educational Measurement*, *52*(3), 319–338. http://dx.doi.org/10.1111/jedm.12078

Brinkhuis, M. J. S., & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems* (Measurement and Research Department Reports No. 09-01). Arnhem, Netherlands: Cito. https://www.researchgate.net/publication/242357963

Brinkhuis, M. J. S., & Maris, G. (2010). *Adaptive estimation: How to hit a moving target* (Measurement and Research Department Reports No. 10-01). Arnhem, Netherlands: Cito. http://www.cito.nl/onderzoek%20en%20wetenschap/achtergrondinformatie/publicaties/measurement_reports

Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, *4*(4), 379–426. http://dx.doi.org/10.1207/s15516709cog0404_3

Coomans, F., Hofman, A., Brinkhuis, M. J. S., van der Maas, H. L. J., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PLOS ONE*, *11*(5), 1–19. http://dx.doi.org/10.1371/journal.pone.0155149

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. http://dx.doi.org/10.1007/BF02310555

Elo, A. E. (1978). *The rating of chess players, past and present*. London: B. T. Batsford.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, *48*, 377–394. http://dx.doi.org/10.1111/1467-9876.00159

Glickman, M. E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, *28*(6), 673–689. http://dx.doi.org/10.1080/02664760120059219

Groeneveld, C. M. (2014). Implementation of an adaptive training and tracking game in statistics teaching. In M. Kalz & E. Ras (Eds.), *Computer assisted assessment: Research into e-assessment* (Vol. 439, pp. 53–58). Springer.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

http://dx.doi.org/10.1007/978-3-319-08657-6_5

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hofman, A. D., Visser, I., Jansen, B. R. J., & van der Maas, H. L. J. (2015). The balance-scale task revisited: A comparison of statistical models for rule-based and information-integration theories of proportional reasoning. *PLOS ONE*, *10*(10), e0136449. http://dx.doi.org/10.1371/journal.pone.0136449

Hofman, A. D., Visser, I., Jansen, B. R. J., Marsman, M., & van der Maas, H. L. J. (2017). Fast and slow strategies in multiplication. Preprint. http://dx.doi.org/10.17605/OSF.IO/AW3QQ

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. (A. Parsons & S. Milgram, Trans.). New York: Basic Books.

Jansen, B. R. J., Louwerse, J., Straatemeier, M., van der Ven, S. H. G., Klinkenberg, S., & van der Maas, H. L. J. (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

Klinkenberg, S. (2014). High speed high stakes scoring rule. In M. Kalz & E. Ras (Eds.), *Computer assisted assessment: Research into e-assessment* (Vol. 439, pp. 114–126). Springer. http://dx.doi.org/10.1007/978-3-319-08657-6_11

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

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. http://dx.doi.org/10.1126/science.1248506

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

Maris, G., & van der Maas, H. L. J. (2012). Speed–accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633. http://dx.doi.org/10.1007/s11336-012-9288-y

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. http://dx.doi.org/10.1207/s15366359mea0204_1

Nižnan, J., Pelánek, R., & Řihák, J. (2015). Student models for prior knowledge estimation. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 109–116). International Educational Data Mining Society. http://educationaldatamining.org/EDM2015

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32. http://dx.doi.org/10.1016/j.intell.2011.11.002

Pelánek, R. (2014). Application of time decay functions and the Elo system in student modeling. In J. Stamper et al. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July 2014, London, UK (pp. 21–27). International Educational Data Mining Society. http://educationaldatamining.org/EDM2014

Pelánek, R., Papoušek, J., Řihák, J., Stanislav, V., & Nižnan, J. (2017). Elo-based learner modeling for the adaptive practice of facts. *User Modeling and User-Adapted Interaction*, *27*(1), 89–118.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute of Educational Research.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Savi, A. O., van der Maas, H. L. J., & Maris, G. K. J. (2015). Navigating massive open online courses. *Science*, *347*(6225), 958. http://dx.doi.org/10.1126/science.347.6225.958

Savi, A. O., Ruijs, N. M., Maris, G. K. J., & van der Maas, H. L. J. (2018). Delaying access to a problem-skipping option increases effortful practice: Application of an a/b test in large-scale online learning. *Computers & Education*, *119*, 84–94. http://dx.doi.org/10.1016/j.compedu.2017.12.008

Savi, A. O., Williams, J. J., Maris, G., & van der Maas, H. L. J. (2017, February 27). The role of A/B tests in the study of large-scale online learning. Preprint. http://dx.doi.org/10.17605/OSF.IO/83JSG

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520. http://dx.doi.org/10.1016/0010-0285(76)90016-5

Sonas, J. (2005). Chessmetrics formulas: Chessmetrics rating as "a weighted and padded simultaneous performance rating." http://www.chessmetrics.com/cm/CM2/Formulas.asp

Straatemeier, M. (2014, April 25). *Math Garden: A new educational and scientific instrument* (PhD thesis). University of Amsterdam. http://hdl.handle.net/11245/1.417091

van den Bergh, M., Hofman, A. D., Schmittmann, V. D., & van der Maas, H. L. J. (2015). Tracing the development of typewriting skills in an adaptive e-learning environment. *Perceptual and Motor Skills*, *121*(3), 727–745.

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2002). *Computerized adaptive testing: Theory and practice*. Netherlands: Springer. http://dx.doi.org/10.1007/0-306-47531-6

van der Ven, S. H. G., Straatemeier, M., Jansen, B. R. J., Klinkenberg, S., & van der Maas, H. L. J. (2015). Learning multiplication: An integrated analysis of the multiplication ability of primary school children and the difficulty of single digit and multidigit multiplication problems. *Learning and Individual Differences*, *43*, 48–62. http://dx.doi.org/10.1016/j.lindif.2015.08.013

Veldkamp, B. P., Matteucci, M., & Eggen, T. J. H. M. (2011). Computerized adaptive testing in computer assisted learning? In S. De Wannemacker, G. Clarebout, & P. De Causmaecker (Eds.), *Interdisciplinary approaches to adaptive learning: A look at the neighbours* (Vol. 126, pp. 28–39). Springer. http://dx.doi.org/10.1007/978-3-642-20074-8_3

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. http://dx.doi.org/10.1111/bmsp.12054

Wauters, K., Desmet, P., & Van den Noortgate, W. (2010). Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, *26*(6), 549–562. http://dx.doi.org/10.1111/j.1365-2729.2010.00368.x

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. http://dx.doi.org/10.1016/0001-6918(77)90012-9

ISSN 1929-7750 (online). The Journal of Learning Analytics works under a Creative Commons License, Attribution - NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0)