UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Psychometrics of an Elo-based large-scale online learning system

Vermeiren, Hanke; Kruis, Joost; Bolsinova, Maria; van der Maas, Han L.J.; Hofman, Abe D.

DOI
10.1016/j.caeai.2025.100376

Publication date
2025

Document Version
Final published version

Published in
Computers and Education: Artificial Intelligence

License
CC BY-NC-ND

Link to publication

Citation for published version (APA):
Vermeiren, H., Kruis, J., Bolsinova, M., van der Maas, H. L. J., & Hofman, A. D. (2025). Psychometrics of an Elo-based large-scale online learning system. *Computers and Education: Artificial Intelligence*, *8*, Article 100376. https://doi.org/10.1016/j.caeai.2025.100376

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:23 Jan 2026

Computers and Education: Artificial Intelligence 8 (2025) 100376

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Psychometrics of an Elo-based large-scale online learning system

Hanke Vermeiren [a,*], Joost Kruis [b], Maria Bolsinova [c], Han L.J. van der Maas [d], Abe D. Hofman [d,e]

[a] Faculty of Psychology and Educational Sciences, and Imec research group Itec, KU Leuven, Kortrijk, Belgium
[b] Cito Institute for Educational Measurement, Arnhem, the Netherlands
[c] Methodology and Statistics, Tilburg University, Tilburg, the Netherlands
[d] Psychological Methods, University of Amsterdam, Amsterdam, the Netherlands
[e] Prowise, Amsterdam, the Netherlands

A R T I C L E I N F O

Keywords:
Architectures for educational technology system
Computer adaptive practice
Elo rating system

A B S T R A C T

The Elo rating system (ERS), an intuitive and computationally efficient algorithm, offers a means to effectively update estimates of item difficulties and learner abilities as they evolve. This method proves to be highly advantageous in online learning environments. Computerized adaptive practice (CAP) endeavors to present learners with items that are well-suited to their individual ability levels, with the ultimate goal of enhancing motivation and optimizing learning outcomes. The objective of this paper is to outline common challenges that arise in an Elo-based CAP system and to present the psychometric enhancements implemented in the Prowise Learn environments to address these concerns. More specifically, we focus on three main aspects; 1) the development of a new scoring rule balancing response time and accuracy, 2) a way to fix the item scale to deal with item drift, and 3) an improved adaptive K-factor algorithm to speed up convergence in estimation. Using data from the Prowise Learn environment, analyses were done to illustrate the effect of the enhancements. Results show that these enhancements result in more dynamic tracking of the ratings, solve the issue of item drift, and capture the speed-accuracy trade-off more accurately.

1. Introduction

Over the last two decades, educational technology (edtech) has become a global phenomenon as a result of a series of innovations such as learning analytics, massive open online courses (MOOCs), and E-learning (Weller, 2018; Zhang & Aslan, 2021; Kew & Tasir, 2022). These innovations have collectively contributed to the widespread adoption and integration of technology in education, revolutionizing the way knowledge is acquired and disseminated. In this changing landscape, learning is no longer confined by temporal or spatial limitations, making education increasingly accessible and providing learners more autonomy and independence (Adedoyin & Soykan, 2023).

One of the most intriguing and exciting aspects of some edtech tools is that they offer the opportunity to customize learning activities according to the needs of the learner, making it a more effective and efficient process (Martin et al., 2020; Bernacki et al., 2021). The presence of individual differences in the classroom makes it impossible to develop an instructional approach that fits all learners. Personalizing the learning process fosters equal opportunities for learners with different skill levels and backgrounds, a feat unattainable in a classic classroom setting where a one-size-fits-all approach dominates. While the idea is not new (Skinner, 1958; Pressey, 1927), personalization has received considerable attention in recent years (Li & Wong, 2023; Gligorea et al., 2023). This can partly be attributed to the notion of the 2-sigma effect (Bloom, 1984), which suggests that tailoring education to the needs of individual learners can significantly enhance learning outcomes. Additionally, the surge in technological advancements allowed for more efficient ways of personalizing the learning process.

A notable challenge faced in online learning environments (OLEs) is the phenomenon of high drop-out rates. Various studies have reported that in online courses completion rates are as low as 10% (Gütl et al., 2014; Eriksson et al., 2017). Numerous factors have been identified as possible causes, including demographic variations, academic experience and behavioral and psychological factors such as low motivation (Lee

* Corresponding author.
E-mail addresses: hanke.vermeiren@kuleuven.be (H. Vermeiren), joost.kruis@cito.nl (J. Kruis), m.a.bolsinova@tilburguniversity.edu (M. Bolsinova), h.l.j.vandermaas@uva.nl (HL.J. van der Maas), a.d.hofman@uva.nl (A.D. Hofman).

https://doi.org/10.1016/j.caeai.2025.100376
Received 6 November 2024; Received in revised form 21 January 2025; Accepted 24 January 2025
Available online 4 February 2025
2666-920X/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

& Choi, 2011; Shaikh & Asif, 2022; Hachey et al., 2023; Karsen et al., 2023). Among these, lack of engagement has emerged as a prominent cause of drop-out (Schaeffer & Konetes, 2010; Wang et al., 2023b). As such, developing a successful OLE brings with it the challenge to keep the learner engaged and motivated. A key element in achieving this goal is to ensure that learners are provided with tasks that challenge their abilities, while also fostering a sense of capability (Jeno et al., 2023; Guay, 2022; Bureau et al., 2022; Kaya & Ercag, 2023). Adaptivity can be a powerful tool in that respect, as it allows us to personalize the OLE to the needs of the learner. For example, by employing adaptive item selection, learners are provided with items that align with their specific ability levels, thus ensuring an optimal level of difficulty (ten Broeke et al., 2021; Wang et al., 2023a).

1.1. Computerized adaptive practice framework

Implementing adaptive item selection in an online learning system allows addressing individual differences in ability levels and learning paths. Students are no longer guided through the learning content in the same manner, but instead are presented with items tailored to their own experience. The value of adaptive item selection in a learning context finds support in various educational theories. Vygotsky's theoretical framework concerning the zone of proximal development (ZPD) posits that presenting learners with tasks slightly beyond their current level of ability not only facilitates optimal learning but also helps maintain an equilibrium between challenge and frustration (Vygotsky & Cole, 1978).

This balance of challenge and affect is also underscored by the flow theory by Csikszentmihalyi (2013) and the self-determination theory (SDT) by Deci and Ryan (2004, 2013). To cultivate a sense of competence in learners, it is imperative that learning items neither pose excessive difficulty nor lack challenge. This assertion also aligns with expertise learning principles and the concept of deliberate practice proposed by Ericsson et al. (1993).

To implement adaptive item selection, a certain student model is needed. Student models form the foundation for personalization in OLEs, since they are tasked with representing the student as accurate as possible, allowing for the identification and comprehension of their individual needs (Pelánek, 2017; Eglington & Pavlik, 2023). In the context of adaptive item selection, the primary emphasis of student models lies in monitoring the ability or skill levels of individuals (Klinkenberg et al., 2011). The estimation of this ability or skill level is generally based on their pattern of correct and incorrect responses on the previously presented items. However, it is worth noting that these models can be tailored to measure various learner characteristics, extending beyond the sole focus on tracking abilities (Chrysafiadi & Virvou, 2013).

Throughout the years, numerous student models have been developed to measure ability levels (Pelánek, 2017). Notable examples include Bayesian knowledge tracing (BKT) (Corbett & Anderson, 1994), performance factor analysis (PFA) (Pavlik et al., 2009) and different models in the item response theory (IRT) (Van der Linden & Hambleton, 2016) framework. However, since most of these methods encounter limitations when implemented in large scale learning environments, rating systems, especially the Elo rating system (ERS) (Elo, 1978) has known widespread application (Kandemir et al., 2024; Klinkenberg et al., 2011; Papousek et al., 2014; Pelánek, 2016; Mangaroska et al., 2019; Ooge et al., 2024; Zhang et al., 2023).

One implementation of adaptive item selection is made in the Computerized adaptive practice (CAP) framework (Klinkenberg et al., 2011). CAP draws inspiration from computerized adaptive testing (CAT) (Meijer & Nering, 1999) where the aim is to improve the efficiency of testing by strategically selecting the most informative items for an individual. Selecting the most informative items at each point in the assessment allows for an equally accurate ability estimation as obtained through standardized testing, albeit in a more time-efficient manner. In contrast, CAP employs adaptive item selection to foster learner motivation and encourage ongoing engagement with the learning environment.

Adaptive item selection is a crucial aspect in both CAT and CAP. However, it is important to acknowledge that the application of the CAT framework to a learning environment is not a straightforward process due to various reasons (Klinkenberg et al., 2011). First, to select the most informative items, it is imperative that item difficulties are known. Traditionally, in the CAT framework, this is achieved by calibrating the item bank using standardized tests (Kingsbury, 2009). In the context of learning environments, this becomes unattainable due to the extensive number of items needed (in our case ~160,000 items). Second, CAT operates optimally when items are presented with a 50% probability of being answered correct (Eggen & Verschoor, 2006). However, within a learning setting, this probability is considered too low to foster motivation. For this reason, it is customary in CAP to set the success rate at .75 or higher. Third, in CAT, it is assumed that there is no learning. This assumption is problematic in dynamic learning environments where the ability of individuals is expected to evolve and change over time. As such, a CAP framework requires an estimation technique that can be used for both item and ability parameters, and is able to track changes in the parameters over time.

1.2. On-the-fly estimation of item and learner parameters

The ERS shares its underlying measurement model with the Rasch model (Rasch, 1960), which serves as a student model relating observed data to the measurement of underlying (latent) person abilities. In the Rasch model, the probability of a correct observed response is modeled by a logistic function based on the difference between the person ability (θ_p) and the item difficulty (δ_i):

$$\mathbb{P}\left(X_{pi} = x_{pi} \mid \theta_p, \delta_i\right) = \frac{\exp\left(\theta_p - \delta_i\right)}{1 + \exp\left(\theta_p - \delta_i\right)}. \tag{1}$$

In a learning context, the system needs to track the ability of a learner and adjust subsequent item selection accordingly, taking into account newly acquired information. On the item side, new data provides information about the relative difficulty of the items, resulting in a reordering of items in the item bank. This self-learning element allows to launch the OLE while some key model parameters will be tuned while data is being collected.

Hence, it is essential that the model parameters (learner ability and item difficulty) are updated as data comes in. As first introduced by Klinkenberg et al. (2011) and further described by Pelánek (2016) the ERS provides a way of achieving that objective. The ERS was originally introduced to track the rating of chess players (player vs player system) (Elo, 1978). Klinkenberg et al. (2011) extended the ERS system to a player vs item system, using the ERS updating rule on both the learner and the items side:

$$\begin{aligned} \theta_p^t &= \theta_p^{t-1} + \mathcal{K}(S_{pi} - \mathbb{E}(S_{pi})), \\ \delta_i^t &= \delta_i^{t-1} - \mathcal{K}(S_{pi} - \mathbb{E}(S_{pi})), \end{aligned} \tag{2}$$

where S is the observed score (correct or incorrect), $\mathbb{E}(S)$ is the expected score that follows from the Rasch model, and K is the K-factor that determines the weight of the new data on the estimation of the model parameters (modeling the bias-variance trade-off). Since the ERS allows updating item parameters, there is no need for pre-calibration of the item bank, which due to the large amount of items needed in a learning environment would become unattainable. A personalized system reacts to what it observes. By relying on a combination of the ERS and a scoring rule, the CAP framework is able to estimate the parameters on the fly and provide adaptivity in terms of item selection. The application of established measurement frameworks, such as Item Response Theory (IRT), enables the creation of sophisticated learning environments. These tools provide a reliable means of explaining and interpreting measurement outcomes, a feature that is not always guaranteed with certain artificial intelligence techniques. (Gligorea et al., 2023; Fiok et al., 2022).

(a) Example of a personal landing page. (b) Example of an item from the addition game.

Fig. 1. Visualization of the Math Garden learning environment.

2. Aim of the paper

While the ERS has been successfully implemented in OLEs, its basic form is not suited to deal with several obstacles. Klinkenberg et al. (2011) presented extensions to the basic ERS algorithm as implemented in a CAP framework. We believe that in the spirit of open science Edtech companies should publish their algorithms allowing for transparency and accountability. As such, the general motivation of this paper is to present further methodological advances of an ERS based adaptive learning algorithm that allows to solve common problems encountered in an ERS system based OLE. More specific, we present three important updates: (1) a student model that follows from the new fixed penalty scoring rule; (2) a way of ensuring the identification of the item bank in systems where item parameters are updated while data come in and (3) a way of defining rating uncertainty, allowing person parameters to change when they should. These implementations aim to enhance the performance and the accuracy of the ERS. Accurate ratings are essential for the process of adaptive item selection, as they ensure students are presented with items of optimal difficulty. Phenomena such as item drift can affect the accuracy of ratings, significantly influencing item selection. By incorporating a well-defined speed-accuracy model and considering uncertainty, our approach ensures that the system maximizes the utility of the available data. This leads to faster convergence, which is particularly crucial at the beginning of the rating estimation process to prevent student drop-out.

2.1. Prowise learn

A successful implementation of the CAP framework are the Prowise Learn learning environment (Klinkenberg et al., 2011). Math Garden is the first of three different adaptive learning platforms launched in 2008 (Fig. 1). Next to the Math Garden, the Language Sea (application for Dutch spelling and grammar learning (Vermeiren et al.)) and Words and Birds (application for Dutch to English learning) (Mulder et al., 2021) were launched. Originally, these systems were developed to collect learning data for fundamental research into cognitive (mathematical) development in an ecologically valid way (Straatemeier, 2014). The great popularity of the system in the Netherlands resulted in spin-off company (Oefenweb.nl, later Prowise Learn) aimed to further develop these systems. In total, these systems currently (2024-01-01) include 65 active games, consisting of around 160.000 active items. The learning systems have about 40.000 unique players per week and collect about 1 million daily responses.

As the primary objective of an adaptive OLE is the development of skills and abilities, it is crucial to engage in frequent and rigorous practice to achieve proficiency. In order to maintain and enhance their current level of ability, learners need to frequently engage with the learning environment. To foster regular participation, all Prowise Learn learning environments employ gamification strategies, which have been proven to prompt regular practice and support the creation of learning habits (Debeer et al., 2021; Chen & Liang, 2022; Zeybek & Saygı, 2024). The use of game-based elements is already apparent when a student opens the learning environment. Prowise Learn environments are characterized by their unique landing pages, each tailored to a specific theme (in the case of Math Garden, this is a garden) where each game is represented by an object according to this theme (plants). Displayed on their personal landing page are games that are recommended for the learners, encompassing advice games or games intended to fulfill learning objectives (Brinkhuis et al., 2020). Furthermore, the page includes new games or games handpicked by their teachers. Besides a student's personal landing page, the learning environment consists of subgardens that are organized according to subdomains such as basic operations (e.g. multiplication), patterns (e.g. completing sequences), and measurement (e.g., clock reading).

Upon selecting a game, learners will be presented with a sequence of 10 exercises, after which they will be redirected to the landing page. Before every game, learners can change the difficulty level they are playing at. The Prowise Learn environments provide three difficulty settings: easy, moderate and difficult (Jansen et al., 2016). These settings correspond to different probabilities of answering the item correctly, with the easy setting having a 90% probability of success, the moderate setting having a 75% probability, and the difficult setting having a 60% probability. It should be noted that the implementation of higher success rates inevitably leads to a loss of valuable information concerning the items. By incorporating response times, the Prowise learning environments aims to mitigate this concern (Klinkenberg et al., 2011).

While learners are practicing their skills by playing various games, coins can be collected. The number of coins a learner receives depends on several things: 1) how long it takes them to answer an item (the faster they are, the more coins they receive, that is if they answer correctly), 2) what difficulty level they are playing on (the higher the difficulty level, the more coins), and 3) what game they are playing (learners are rewarded for playing advice games). In case children absolutely do not know the answer, they also are given the option to select a question mark button, (Savi et al., 2018) in which case they do not receive coins. To keep track of their progress, a learner's ability can be found both on their growing card and in their gardens. Ability levels are represented by a score from 0 to 1000 (q-scores) that represent how well learners do with respect to their peers (e.g. someone at the end of grade 3 should have a score of around 300).

The Prowise Learn environments implement governing strategies both on the item and the games level. Governing strategies are meant to guide the students through the learning environment without teacher supervision. Guided by the national curriculum, and depending on the grade of the student, initially only a subset of games is available. Once a student reaches a certain level in these games, the more advanced domains can be unlocked. The main principles to guide learners to different domains is (1) the suitability of the game for the specific grade, (2) the amount of practice on the specific game, (3) the relative performance of the learner on this game compared to other domains, and lastly (4) it is ensured that domains are not recycled each day (Brinkhuis et al., 2020). This is done to prevent that learners are too often confronted with their weakest domains, which might affect their motivation to continue practicing. These governing strategies ensure that children play the games on which they can show progress, either because scores are lower compared to other learners or because they score lower on this domain given their scores on other domains.

3. Updates to the student model and the ERS

In the last ten years have not only the number of games and disciplines that are offered in Prowise Learn increased, under the hood also several improvements of the adaptive algorithm have taken place. In this section, we will discuss three of those updates. First, we discuss an update to the scoring rule that influences the observed and expected score. Second, we discuss an update to the process by which we update the item parameter estimates. Third, we discuss an update to the calculation of the K-factor values. We first will briefly discuss the initial configuration in the case of the scoring rule and the K-factor (for a more thorough discussion we refer the reader to the paper by Klinkenberg et al. (2011)), next we explain why the development of these three feature was advisable and the solution that was implemented, followed by a discussion of the effects of this change on the system.

3.1. Fixed penalty scoring rule

To keep learners motivated, in CAP items are selected with a success rate higher than 50%. As described in Klinkenberg et al. (2011) response times can then provide valuable additional information about the learner's ability and therefore using this information allows for quicker parameter convergence. However, how accuracy and response time information should be combined, is a long debate in psychological testing (Maris & Van der Maas, 2012; Van der Linden, 2007; Vandierendonck, 2021). Traditionally, in psychometrics, the focus is on one of the two while the other is ignored. Another option is the use of models that aim to model the speed-accuracy trade-off (Van der Linden, 2007; Ranger & Kuhn, 2012; Molenaar et al., 2015; Rouder et al., 2015; Heathcote & Matzke, 2022). However, these models do not provide information on how the learners should weigh the importance of their response time. In the Prowise Learn systems, this is solved by adopting an explicit scoring rule approach, where learners are informed about how response times influence their scores by the remaining coins that can be won by providing a correct response. The observed score S in (2) is now no longer 0 or 1 but instead calculated based on a scoring rule that includes time. Initially, the high-speed-high-stakes (HSHS) scoring rule as described by Maris and Van der Maas (2012) was implemented. This rule results in high stakes (win or loose) for quick responses, and is given by the function:

$$S^{\star}_{pi} = (2x_{pi} - 1)(d_i - t_{pi}). \tag{3}$$

It can be seen in Equation (3) that $S^{\star}_{pi}$, the observed score on item i by person p, is a function of x_{pi}, the accuracy of the response on item i by person p; d_i, the time limit for item i; and t_{pi} the response time. For our convenience the time parameters are scaled, specifically, we set $d_i = 1$, and let t_{pi} fall within the range $[0, 1]$, such that it represents the proportion of the time limit on item i taken by person p to give a response. Fig. 2 gives a visual representation of this scoring rule.

Naturally, the goal of this scoring rule was to prevent fast guessing and stimulate careful consideration of the response instead. In order to stimulate children to behave according to this scoring rule we introduced a system where within a session an amount of digital currency (coins) was rewarded relative to the observed score, e.g., quick correct responses would result in high gain, whereas a quick incorrect response would result in a high loss. At the end of a session, the number of accumulated coins in this session would be awarded to the child, and they could then use this currency to buy prizes from a digital prize cabinet.

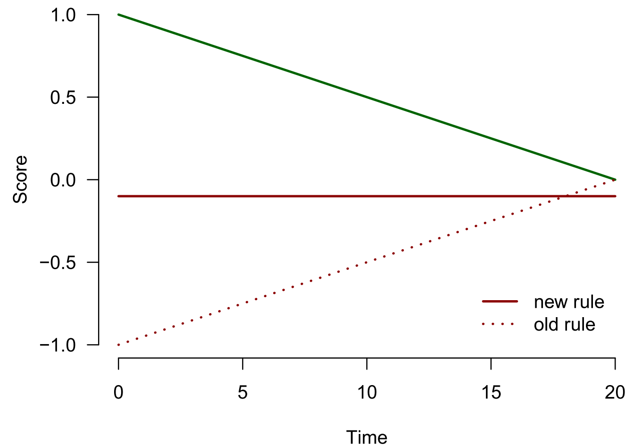

Fig. 2. Visualization of how the HSHS and the FP scoring rules relate to each other for a deadline of 20 seconds. For the FP scoring rule a fixed penalty of 2 coins is implemented.

From this scoring rule a measurement model can be derived if we assume that $\sum_p S^{\star}_{pi}$ is a sufficient statistic for the item difficulty, $\sum_i S^{\star}_{pi}$ a sufficient statistic for person ability and item responses are conditionally independent. Given these assumptions, the joint density for response time and response accuracy can be formulated, from which the expected score for the HSHS rule can be derived (Maris & Van der Maas, 2012):

$$\mathbb{E}(S^{\star}_{pi}) = \frac{\exp(2(\theta_p - \delta_i)) + 1}{\exp(2(\theta_p - \delta_i)) - 1} - \frac{1}{\theta_p - \delta_i}. \tag{4}$$

From an educational standpoint, it might be difficult to justify the high penalty that occurs for quick incorrect responses. Specifically, we assumed that because the observed score is directly tied to the reward learners received in the form of coins, they will behave according to the scoring rule. However, we received negative feedback by schools and teachers, who informed us that students had an aversion for the high penalty for fast incorrect answers. Literature corroborates the undesired effect of negative feedback on learning and motivation (Van Duijvenvoorde et al., 2008; Mercer & Gulseren, 2024). Furthermore, as can be seen in Fig. 4 the response time data for incorrect responses under the old rule does not fit the model.

Hence, in 2019 we developed a new scoring rule, the fixed-penalty (FP) scoring rule, which introduces the fixed domain penalty $[\rho_g \in [0, 1]]$ and has the following form:

$$S_{pi} = x_{pi}\left(d_i - t_{pi}\right) - \left(1 - x_{pi}\right)\rho_g \tag{5}$$

Fig. 2 illustrates this new scoring rule with a fixed penalty for negative scores. A comparison of the initial HSHS scoring rule with score $S^{\star}_{pi}$ and the FP scoring rule with score S_{pi} shows that for the correct response, both scoring rules result in the same score (d_i - t_{pi}). However, whereas an incorrect response in the HSHS scoring rule resulted in a score of $t_{ij} - d_i$, in the new scoring rule the score for an incorrect response is no longer dependent on the response time, but instead constant and equal to $-\rho_g$, the fixed penalty for domain g. This penalty parameter is set for each domain and determines the proportion of the scaled time limit, that is given as a penalty for an incorrect response.

The new FP rule gives the following formula for the expected score:

$$\mathbb{E}(S_{pi}) = \frac{-\rho_g \vartheta^2_{pi}\exp(-\rho_g\vartheta_{pi}) + (\vartheta_{pi} - 1)\exp(\vartheta_{pi}) + 1}{\vartheta_{pi}(\exp(\vartheta_{pi}) - 1 + \vartheta_{pi}\exp(-\rho_g\vartheta_{pi}))} \tag{6}$$

with $\vartheta_{pi} = \theta_p - \delta_i$, the difference of the ability of the person (θ_p) and the difficulty of the item (δ_i). For the derivation, see Appendix A.

In practice, we generally have set $\rho_g = .1$, such that the score for an incorrect response becomes $-.1$ and the number of coins a learner would lose from their session total for an incorrect response would also

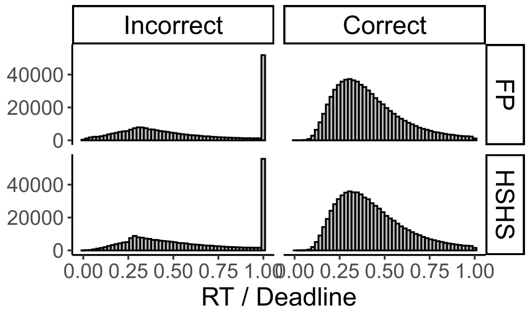

Fig. 3. Distribution of the reaction times for correct and incorrect answers for both rules.

be equal to 10% of the maximum reward for a correct response. To illustrate the difference between the scoring rules, we take the situation in which the time limit of an item was 20 seconds, and the learner gave an incorrect response after 8 seconds. For the initial HSHS scoring rule this would result in a score of $-.6$ and they would lose $.6 \times 20 = 12$ coins from their session total, for the FP scoring rule the learner would however only loose $.1 \times 20 = 2$ coins from their session total. Hence, only in cases where $x_{pi} = 0$ and $t_{pi}/d_i > (1-\rho_g)$ will we find that $S^{\star}_{pi} > S_{pi}$.

Naturally, the goal of this scoring rule is still to stimulate careful consideration of the response instead. However, by not penalizing incorrect responses inversely to response time, we aimed to ensure that children do not become demotivated by incorrect responses, as these are an integral part of any learning process. One could argue against this new rule by pointing out that not punishing fast incorrect responses, might incentivize fast guessing. However, Fig. 3 shows that no difference is found in the RT distribution for the HSHS and the FP rule. Implementing a fixed penalty thus does not prompt learners to give faster incorrect responses. For this figure, random samples from person item interactions in the addition game were selected. Both for the old and the new rule 1.000.000 responses were sampled resulting in data from 100854 unique players.

Secondly, Fig. 4 shows that the FP rule provides a better fit of the response times compared to the HSHS scoring rule. For Fig. 4 two datasets from the addition game were used, one from when the HSHS rule was in place, and another more recent dataset under the new FP rule. Data was selected for items of all difficulty settings. Model expectations (black lines) were plotted to examine the fit of the obtained data (colored lines) to the models of the scoring rule. It is clear that for the old rule, especially the data for incorrect responses deviates from the expected pattern. According to the model, response time should increase with ϑ_{pi}. In other words, for incorrect answers, the model expects that a higher ϑ_{pi} results in more deliberate answers and thus slower response times. However, the data indicates the opposite pattern, with higher ϑ_{pi} resulting in faster response times.

Data from the new rule shows a better fit for incorrect answers and a slightly better fit for the correct answers to the model. For both the old and the new rule, in general those with a higher ϑ_{pi} have a faster response time for incorrect responses. In other words, those with a higher probability of answering correct display faster reaction times. Note that for the old rule, the effect of ability on reaction time is in the opposite direction than what the model predicts. The HSHS rule expects those with higher ϑ_{pi} to longer consider their answers when unsure about the correct answer. For the FP rule, the model predicts no effect of ϑ_{pi} on reaction time. While the data shows some divergence from this prediction, the FP rule shows a better fit to the data than the HSHS rule. A possible reason for the phenomenon that those with higher ϑ_{pi} give faster incorrect answers, is that for those students, incorrect answers result from careless mistakes rather than a lack of comprehension.

The fact that the distribution of reaction times for incorrect responses remains unchanged, along with improved fit of the new rule to the data, provides additional justification for the new rule which was initially adopted for pedagogical objectives.

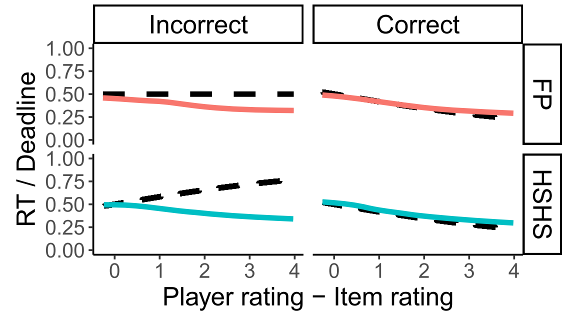

Fig. 4. Fit of the data gathered under the scoring rules to the expected reaction time for correct and incorrect answers according to the models of the scoring rules. Black dotted lines represent the model predictions.

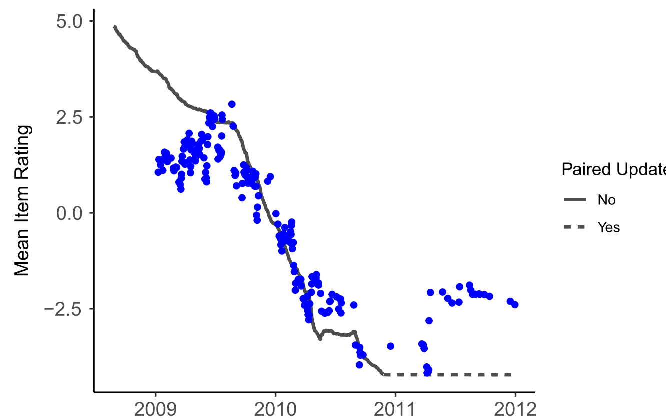

Fig. 5. Visualization of the effect of the paired updates algorithm on item drift. Black line indicates the evolution of the average item rating. Blue dots indicate the evolution of one single user.

3.2. Fixing drift in the item parameters

To bypass item calibration, item difficulty is estimated in real time as data comes in. However, with the implementation of the initial version of the ERS we encountered problems of rating drift. In hindsight, this is to be expected when the ERS is implemented in learning systems. As in IRT scales, some kind of anchoring is required (Sinharay et al., 2011). As can be seen in (2), in applying the ERS, when learners answer an item correct their ability estimates increase by the same amount that the item difficulty estimate decreases. As such, when learners learn, they take rating points from items, resulting in decreasing items ratings over time. Fig. 5 show this drift in one of the Prowise Learn games. In a period of two years, the average item rating drops from 5 to -4. This item drift results in model identification issues and thereby in difficulties of parameter interpretation.

While item drift does not impact the comparison and ranking of the learners and items at one point in time, including adaptive item selection, it poses a problem for comparison of ratings over time. Furthermore, item bank drift can have negative effects on item selection for learners that have not been active in the system for a considerable period. The item bank has deviated from its original state, resulting in the misrepresentation of item difficulty ratings, with items appearing to be less challenging than they truly are. As a result, when these learners again enter the system, they will be confronted with items that are too difficult given their ability levels. For example, the blue dots refer to the rating of a single learner. In this case, the learner's rating decrease follows the average item ratings, hence it appears the learner gets worse over time.

To solve the problem of rating drift we introduced paired item updates. Instead of updating an item rating after each response, we update items in pairs according to the following rule.

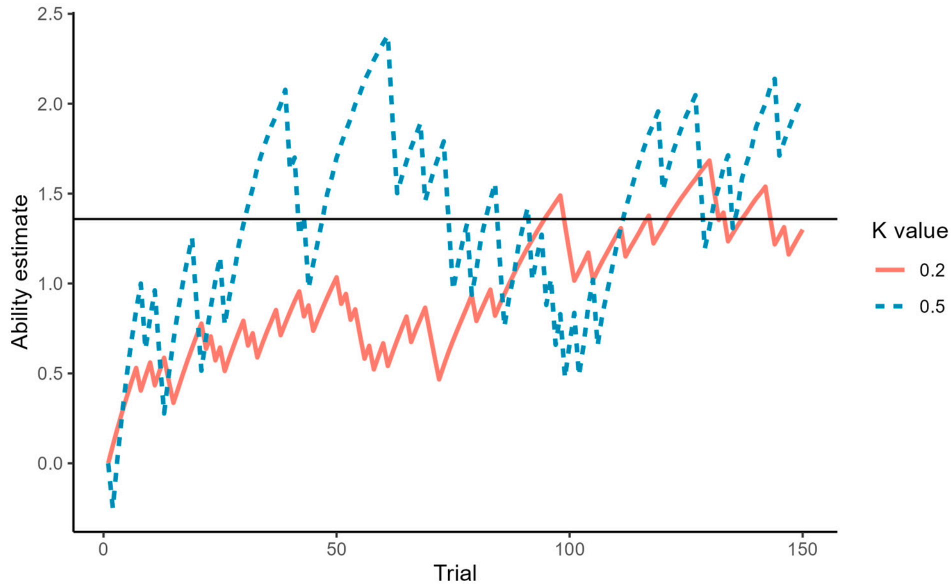

Fig. 6. Trade-off between fast convergence and stable estimates for low and high K values.

$$\text{if } \operatorname{sgn}\left(S^t_{pi} - \mathbb{E}(S^t_{pi})\right) \neq \operatorname{sgn}\left(S^{t-1}_{pj} - \mathbb{E}(S^{t-1}_{pj})\right) \rightarrow \begin{cases} \delta^t_i = \delta^{t-1}_i + \kappa \\ \delta^t_j = \delta^{t-1}_j - \kappa \end{cases} \tag{7}$$
$$\kappa = \frac{1}{2}\left(\mathcal{K}^t_i\left(S^t_{pi} - \mathbb{E}(S^t_{pi})\right) - \mathcal{K}^{t-1}_j\left(S^{t-1}_{pj} - \mathbb{E}\left(S^{t-1}_{pj}\right)\right)\right)$$

where the change in rating κ is a function of the difference between the observed and expected score on two items i and j administered in succession, each weighed by their respective K-factors.

If a learner wins from an item (taking some rating points of that item) we only update the item if the same learner also lost from the previous item within the same session, or vice versa. By keeping this time window restricted to successive items in the same session, we ensure that these changes can be made under the assumption of stable true ability. If there is a match, we update both items with the average change, one upward and one downward. This approach fixes the average item rating at each moment in time at the cost of not always updating the item ratings. In Prowise Learn in about 35% of the responses, the rating gets updated, however due to the large scale of the learning environment this is not problematic. By fixing the item bank in this way, it allows for a direct interpretation of changes in learner ratings, caused by learning (or forgetting). The effect of implementing paired item updates is apparent in Fig. 5, as from 2011 on, item ratings stabilize and the learner's ability rating increase again.

3.3. Ensuring parameter changes, when they are needed

The K-factor ($\mathcal{K}$) is a critical parameter in adaptive algorithms that determines the weight of parameter changes. Its importance lies in the fact that it regulates the extent to which the algorithm adapts to changing data. A small value of $\mathcal{K}$ implies slow adaptation, while a large value of $\mathcal{K}$ leads to fast and potentially unstable adaptation. A correct K-factor ensures that the rating approaches to the ability level at an appropriate rate (Fig. 6). Having reliable ratings for a learner in an adaptive system is really important in order to keep the motivation to play high. If a learner continuously gets administered items that are inappropriate for their ability level, they will become bored (underestimated ability) or demotivated (overestimated ability) while playing. When the current estimate of a learner's ability is no longer correct and the K-factor is too low, the rating will move in the direction of the true ability level very slowly, and the learner will not receive appropriate feedback about their progress. On the other hand, if the K-factor is too high, the rating will overshoot. Therefore, choosing an appropriate K-factor is critical to optimizing algorithm performance. This is especially important when a new learner enters the system and there is little or no data available about their ability level, which is known as the cold-start problem. The initial rating assigned to a learner is generally not representative of their ability level, and in this situation, the K-factor becomes particularly important to ensure that the learner is as soon as possible presented with appropriate items.

In the initial configuration of the Prowise Learn environments, the K-factor values for a learner p and item i were specified in the following way (Klinkenberg et al., 2011):

$$\mathcal{K}_p = \mathcal{K}\left(1 + \mathcal{K}^+_i \mathcal{U}_i - \mathcal{K}^-_p \mathcal{U}_p\right)$$
$$\mathcal{K}_i = \mathcal{K}\left(1 + \mathcal{K}^+_p \mathcal{U}_p - \mathcal{K}^-_i \mathcal{U}_i\right), \tag{8}$$

with $\mathcal{K}$ the default value if there is no uncertainty, $\mathcal{U}_p$ and $\mathcal{U}_i$, the rating uncertainty of the learner and item, each weighted by their own (fixed) K-factor, $\mathcal{K}^+$ and $\mathcal{K}^-$.[1]

These weights ensured that changes to the person K-factor are more strongly influenced by person rating uncertainty, and changes to the item K-factor would be more strongly influenced by the item rating uncertainty. Rating uncertainty is a measure of the uncertainty in estimating the student's skill level (or the item difficulty). A high rating uncertainty indicates that the student's ability or item's difficulty is uncertain, and the K-factor should be set to a higher value to allow for more significant changes in the rating as the administration of items progresses. In contrast, a low rating uncertainty indicates that there is more certainty about the student's ability or item's difficulty, and the K-factor should be set to a lower value to avoid over-fitting the model to the student's performance on a few items. In the initial configuration of the Prowise Learn environments, rating uncertainty was a function of the number of administrations and the days of not playing (D):

$$\mathcal{U}^t = \mathcal{U}^{t-1} - \frac{1}{40} + \frac{1}{30}D \tag{9}$$

From this expression it can be seen that, with an initial value of $\mathcal{U} = 1$, it takes 40 administrations for uncertainty to go to 0, and it would go back to 1 after not playing for 30 days. For example, if a student has only completed a few exercises, there will be more uncertainty in estimating their skill level, and a higher K-factor would be appropriate.

[1] The initial configuration of the Prowise Learn systems used $\mathcal{K}^+ = 4$ and $\mathcal{K}^- = 0.5$.

As the student completes more items and their skill level becomes more certain, the K-factor can be reduced to prevent over-fitting. However, if a student has not played for a long time, it is still possible that there has been some change in their ability, and hence we want to quickly be able to pick up on this change and get their rating estimate to the 'real' rating as soon as possible. The rating uncertainty function is based on how rating deviation is implemented in the Glicko rating system (Glickman, 1999).

While this implementation of rating uncertainty is certainly intuitive, it has one drawback, namely that the change in rating uncertainty is completely independent of the performance of the learner on the task, and only focuses on the number of administrations. As such, if the start rating of a learner is very far from the true rating, by the time 40 items have been administered the estimated rating can still be far from the true rating. In addition, even if a student is absent for a short period of time, their ability could have changed, which will be reflected in their performance. For example, if a student practices an ability outside the learning environment for two days, the increase in the value of the K-factor might be too small to keep track of the change in their ability.

An approach to rating uncertainty that addresses the limitations of the previous method is to base the rating uncertainty on the difference between observed and expected response behavior. This method takes into account the learner's performance on the task and adjusts the rating uncertainty based on their response behavior and is therefore an indication of the trend in their data. As such, we no longer speak of an uncertainty parameter but a trend parameter T. If a learner's response behavior is consistent with their current estimated rating, this indicates the absence of trend in the data and the trend parameter will get closer to 0, indicating that their true rating is likely to be close to their estimated rating. Conversely, if a learner's response behavior is inconsistent with their estimated rating, this indicates a trend in the ratings and the absolute value of the trend parameter will get closer to 1, indicating that their true rating is likely to be further from the estimated rating. In 2008, we implemented such a formulation of a rating trend in Prowise Learn in the following way:

$$\mathcal{T}^t = (1-\alpha)\mathcal{T}^{t-1} + \alpha \,\mathrm{sgn}\left(S^t_{pi} - \mathbb{E}(S^t_{pi})\right) \tag{10}$$

In this new equation, the hyperparameter α dictates the speed by which $\mathcal{T}^t$ adapts to new information. Furthermore, $\mathcal{T}^t$ now takes values between -1 and 1 representing whether the learner scores more above or below expectation given their current rating estimate. This rating trend value is then plugged in for the calculation of the K-factor in the following way:

$$\begin{aligned} \mathcal{K}^t_p &= \max(\mathcal{K}^-_p, |\mathcal{T}^t_p|^\lambda) + \mathcal{K}_s(n_p \leq 20) \\ \mathcal{K}^t_i &= \max(\mathcal{K}^-_i, |\mathcal{T}^t_i|^\lambda) \end{aligned} \tag{11}$$

In this expression for the K-factor, the hyperparameter λ shrinks the influence of $\mathcal{T}$ on $\mathcal{K}$, and restricts the range in which $\mathcal{K}$ consequently falls with $\mathcal{K}^-_p = .2$ and $\mathcal{K}^-_i = .001$. By adding a small constant ($\mathcal{K}_s$) to the K-factor on the first 20 administrations for a person, this method also addresses the cold start problem more effectively, as it can quickly adjust the rating trend based on the learner's response behavior, and at the same time allows for bigger steps in the initial phase of the rating calibration. To determine the optimal hyperparameter values for the formulas in (10) and (11) an A/B test was set up in several games of Math Garden.

By basing the rating trend, and therefore also the K-factor, on the difference between the observed and expected response behavior, this approach ensures that ability rating will move faster towards the true ability level while also maintaining stable ratings. Fig. 7 and Fig. 8 visualize the functioning of the adaptive K-factor in two different scenarios. In Fig. 7 ability remains relatively constant over time as indicated by the constant alternation between positive and negative updates. As a result, the trend parameter alternates around the same value, keeping the K-factor constant at a low value. In this case .2 as this is the minimal value implemented in the environments for the K-factor of person updates ($\mathcal{K}_p$). In Fig. 8, however, ability undergoes more changes, especially towards the end of the plot. This learner repeatedly achieves positive scores, indicating that the presented items might be too easy for them. The trend parameter rapidly reacts to this trend in the scores by increasing towards a value of 1. As a result of an increase in the trend parameter, the $\mathcal{K}$ value goes up, allowing larger updates to the ability ratings and thus faster convergence to the true rating. For a more in depth analysis on the effectiveness of an adaptive K-factor function in an Elo-based system, we refer to Vermeiren et al.

4. Discussion

The Prowise Learn environment relies on an Elo-based CAP framework to address individual differences of learners, by tailoring the content to each learner's unique skill level, such that children in the same grade or classroom can be presented with very different content. Ensuring the smooth functioning of the underlying system is crucial in order to facilitate an efficient learning process. CAP is aimed at personalizing the learning process by selecting items best suited for the learner based on their skill levels with the goal of optimizing the learning process. Matching the difficulty of items to the ability of the learner is believed to increase learner motivation and engagement. Increased motivation in turn is associated with better learning outcomes, for instance by encouraging continued practice in the learning environment. To achieve this, CAP environments aim to present learners with items that are challenging but still manageable. Specifically, a CAP implements a success rate of approximately .75. However, for this personalized item selection process to function effectively, the system must be able to provide accurate ratings for both the items and the learners. Inaccuracies as caused by some of the problems discussed in this paper could lead to a mismatch between the learner's ability and the item difficulty, undermining the intended motivational benefits of the system. As such, implementing solutions that deal with problems such as item drift and the cold start problem is essential for a learning environment that aims to foster continued engagement.

In this paper, we described several challenges that occur when implementing the ERS in a large scale learning environment, and which may affect the efficiency of the system. We describe psychometrical improvements of the basis ERS algorithm to deal with these challenges: 1) a fixed penalty scoring rule; 2) a way of ensuring the item bank is fixed in a system where item parameters are updated while data come in and 3) a way of ensuring person parameters to change when they should. For each, the initial version is described, as well why a change was needed and how the problems that occurred with the initial versions were solved. In the spirit of open science in education, we give a detailed overview of the psychometrics underlying these innovations.

A scoring rule allows taking into account response time when determining the score for an item. Leveraging reaction time provides more information about the ability of the learner, which is a valuable advantage in adaptive learning environments where we aim for higher success rates. Items with the highest item information are those with a 50% probability of answering the item correctly. In order to engage and motivate learners, we often aim for a success rate of 75% or more, thereby losing valuable information. Incorporating reaction times allows presenting learners with easier items while still obtaining effective measurement. Initially, the prowise learning environments implemented the HSHS rule, which punishes fast incorrect responses, but rewards fast correct responses. The goal of this rule was to minimize guessing. The change of the scoring rule was not implemented to improve the accuracy of the rating system, but rather to directly influence the learner's motivation. From teacher feedback, it became evident that the initial scoring rule had unintended effects on children's motivation due to the large loss of coins when a fast incorrect answer is given. This large penalty created a demotivating experience, affecting the learner's willingness to engage with the material. Hence, the FP penalty scoring rule, while less con-

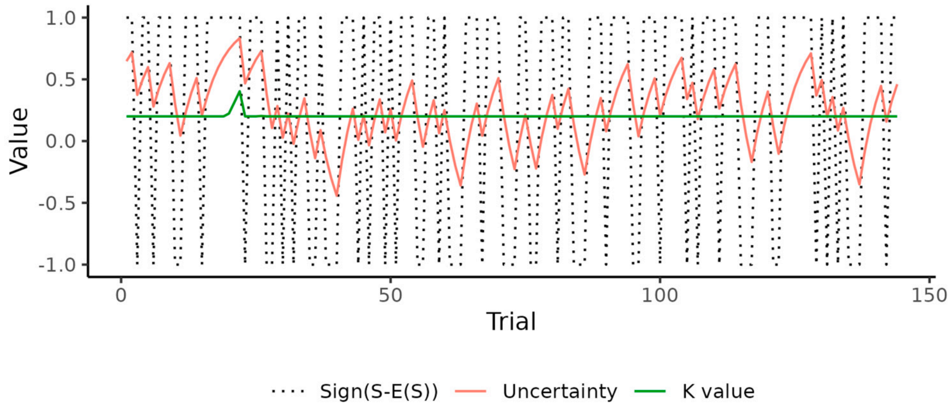

Fig. 7. Visualization of the adaptive K function implemented in the Prowise Learn environments.

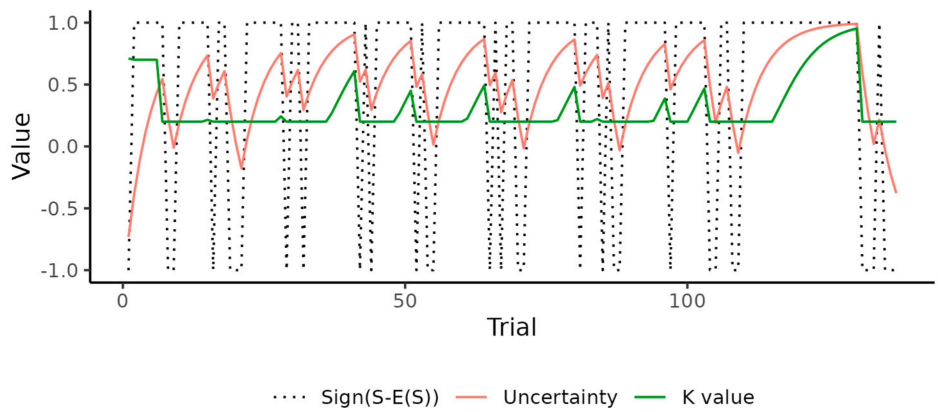

Fig. 8. Visualization of the adaptive K function implemented in the Prowise Learn environments.

venient due to a more complex expected score and item selection rule, is educationally more justifiable. Note that the solution described here is not the only way to incorporate both accuracy and reaction time to calculate an outcome score. The HSHS rule exhibits favorable properties and may yield fewer adverse effects from an educational point of view in older populations. Furthermore, an alternative approach is to separate the rating system and the reward system of the learning environment such that the HSHS rule can be used for the ratings, but learners are no longer confronted with a loss of coins when giving a fast incorrect answer. Other ways of modeling the speed-accuracy trade-off have been proposed (Coomans et al., 2016) and might be modified to be educationally more justified.

As mentioned in the paper, the Prowise Learn environment implements a paired rating update to deal with item drift. This form of item drift will always occur in an Elo-based learning system, as it is inherent to the functioning of the rating system. In general, students improve during their time in the learning system, resulting in an increase of their ratings over time at the expense of the item ratings. We show that on average, item ratings decrease over time if no measures are taken. As mentioned earlier, this can have unintended consequences for learners that have not been active in the learning environment for a while. The item bank has deviated from its original state, resulting in the misrepresentation of item difficulty ratings, with items appearing to be less challenging than they truly are, leading to a mismatch between the difficulty of the item and the current ability of the learner. The use of paired item updates is a simple but effective solution to this challenge. A drawback of this implementation is that the item ratings are updated less frequently. While this is not a problem in a large scale environment as Prowise Learn, in other instance this might pose problems. One option is to implement paired updates using a queue approach. Instead of limiting the pairing to the item that came before it, in this setup items can be paired with any other item from a queue of non-updated items. Note that this can become computationally challenging in the ERS as not only the items need to be remembered by the system but also their update value.

One of the advantages of the CAP framework is that the systems learns the item parameters over time and those do not need to be precalibrated. However, continuously updating item parameters can have undesired effects, such as the item drift we explained earlier. Paired updates solve this general form of item drift inherent to the use of the ERS in a learning system. However, once in a while, a specific issues with the estimation of item difficulties still occurs. One such issue is item clustering. This is caused by learners relying on some strategy they learned to answer a specific kind of item. Employing these strategies typically yields correct responses for only a limited number of items, leading to a consistent decrease in rating for those items as they are repeatedly answered correctly. Conversely, the rating of the other items tends to increase since the strategy results in incorrect responses for these items. As a result, sets of items drift off from the rest of the item bank. Since the other items will become too difficult for some learners, they will never be exposed to them. Several courses of action are possible to tackle this. In the Prowise Learn environment, clustered items are tagged as mirror items. By implementing a certain ring length (the number of items before a new item with the same label is presented) we avoid that learners are presented with similar items in sequence and thus discern a certain strategy that works for these items but not the others.

The last implementation we discussed is that of an adaptive K-factor that gets updated when data comes in. Allowing the K-factor to be rating sensitive, ensures that the size of the parameter updates is appropriate given the current trend in the data. A small value of $\mathcal{K}$ results in slow convergence of the ratings and is ill-suited to deal with changes in ability (which characterizes a learning environment), while a large value of

$\mathcal{K}$ leads to faster convergence but unstable ratings. An adaptive K-factor that can alternate between high and low values ensures stable estimates when the ratings are approaching ability levels, but is still able to adjust rapidly to changing ability levels. As a result, implementing an adaptive K balances flexibility and stability. This ensures that the rating estimate converges to the true rating at an appropriate rate and allows the system to faster escape the cold start for new learners. The cold start problem arises in all student models due to the limited information available when a new learner enters the system. This lack of information can negatively impact learner motivation, as the system may select items that are not optimally matched to the learner's abilities. In an ERS-based system with a fixed value for the K-factor, there is a trade-off between fast convergence, which may result in noisy and less reliable ratings, and slow convergence, which leads to more stable ratings. By implementing a dynamic K-factor approach, it becomes possible to achieve both fast convergence and stable ratings as well as closely track changes in the underlying abilities over time. Reliable ability ratings are the cornerstone of effective adaptation. If learners are continuously presented items that are too easy or too difficulty due to over or underestimation of their abilities, they will get demotivated to continue practicing and drop out.

While the described implementations solve some of the problems we encounter in online learning environments, we recognize that there are still aspects of the ERS that can be improved upon. One of these drawbacks is that the ERS does not provide a measure of error and therefore does not allow a measurement of the estimates accuracy. Urnings, a rating system recently proposed, bypasses this problem by ensuring a known binomial invariant distribution (Bolsinova et al., 2022a, 2022b; Hofman et al., 2020). Since the invariant distribution is known, it becomes possible to calculate confidence intervals for the ratings. As such, if the ratings are to be used in a more high stakes context such as assessment, this approach is more justifiable. Furthermore, Urnings also provides a solution for the variance inflation that occurs in ERS based learning environments. The inflation of the variance of the item ratings is influenced by a variety of things, such as success rate and the K-factor. By implementing a Metropolis-Hastings step, the Urnings algorithm allows remedying this. Research on possible ways this can be solved in the ERS is needed.

In addition, we illustrated that our proposal of a dynamic K-factor is able to track changes in the underlying ability levels. Further research is needed to compare this method with other methods such as the Glicko system as well as investigate its performance in other learning environments. However, given its simple, straightforward implementation we expect that our idea of a dynamic K-factor can benefit many other Elo-based learning environments that now work with fixed K-factors or uncertainty functions where the K-factor is in function of time. While our results show that the new scoring rule better fits the data of the students, further research might look into the effect on the students' motivation to get a better grasp of the effect of the minimal loss students experience when answering an item wrong. One disadvantage of implementing a general scoring rule is that we are again viewing the learners as a homogeneous group. Research is needed to investigate whether groups can be detected that respond differently to the scoring rule. Another possible future direction is expanding the scope of the ERS algorithm to encompass a multidimensional framework. Knowledge pertaining to one particular skill could potentially provide insights into other abilities, even if they are not actively being practiced at the given moment. Furthermore, a multidimensional framework can enhance learning by optimizing how users are navigated through the different games. At the moment, each game and as a consequence all its items are assumed to be unidimensional, implying that only a single skill is required to solve each item. However, the implementation of a within-item multidimensional ERS approach has the potential to yield improved and more accurate measurement (Park et al., 2019a).

This paper demonstrates that it is feasible to build a large scale online learning environment based on an autonomous system that acts independently to a certain degree. That is to say, the system learns the difficulties of items and based on a specified item selection procedure selects an appropriate item for learners without the need for human intervention. This is achieved by implementing an intuitive algorithm, therefore maintaining transparency on how parameters values are obtained. The system delineated here tailors learning experiences to the unique needs of each learner, aiming to increase learning efficiency and effectiveness. By implementing governing strategies, we ensure that learners mostly play games where they can and should develop their ability at that time. Therefore, learners spend most of the time actually playing a diverse set of games. Several psychometric improvements can be implemented that effectively address practical challenges, such as item drift and ensuring accurate estimates when ability levels fluctuate over time. Our results prove that the implemented improvements optimize the system by allowing for more dynamic tracking of the ability levels of the students and preventing the item ratings from drifting. Both of these allow for more accurate item selection, a key factor in providing the students with a learning experience that both challenges and motivates them. All the challenges discussed here are not unique to the Prowise Learn environment, as such we outlined here three implementations that can easily be implemented in any Elo-based learning environment thereby improving its performance.

A well functioning system is not only important for a successful educational application, but also makes the learning environment a valuable asset for scientific endeavors. Data from the Prowise Learn environment have been used in numerous research projects, ranging from improving the underlying ERS algorithm and CAP framework (Coomans et al., 2016; Brinkhuis et al., 2018; Park et al., 2019b; Maris & Van der Maas, 2012; Savi et al., 2018) to research on language and mathematics learning (Hofman et al., 2018c; de Bree et al., 2017; van der Ven et al., 2013, 2015; Hilz et al., 2023), as well as the effectiveness of adaptive item sequencing in terms of learning outcomes (ten Broeke et al., 2021; Jansen et al., 2013, 2016). Moreover, online learning environments as frequently used as Prowise Learn allow collecting extensive longitudinal data (e.g. learning analytics) including log data which make it feasible to study development (Brinkhuis et al., 2015; Braithwaite et al., 2016; Hofman et al., 2018a) as well as propose and assess new theories about the development of intelligence (Van Der Maas et al., 2006; Ou et al., 2019; Hofman et al., 2018b). Log data can also be used to detect strategy use and anomalies, as well as help gain insight in error making (de Mooij et al., 2021). We believe the usefulness of the CAP data stems from the robustness of the underlying measurement model, which is comparable to a Rasch model.

CRediT authorship contribution statement

Hanke Vermeiren: Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Joost Kruis:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Maria Bolsinova:** Writing – review & editing, Conceptualization. **Han L.J. van der Maas:** Writing – review & editing, Conceptualization. **Abe D. Hofman:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Conceptualization.

Statement on open data and ethics

The study was approved by the Ethics Committee of the University of Amsterdam with ID: FMG-3197. The anonymous data for the analyses were selected from schools that gave permission to use the data for research purposes.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: One of the researchers involved in the project is employed one day a week by Prowise, the company that develops the learning software.

Acknowledgements

This work was supported by the Research Foundation Flanders, G0D4122N (Wim Van Den Noortgate), European Research Council advanced grant, 101053880 CASCADE (Han L. J. van der Maas) and Dutch Research Council VI.Veni.221G.056 (Maria Bolsinova).

Appendix A. Derivation expected score FP rule

Just as for the HSHS scoring rule (Maris & Van der Maas, 2012), a model can be derived for the FP rule given some auxiliary assumptions.

First, a person's answers to different items are assumed independent given their ability (conditional independence assumption):

$$\perp\!\!\!\perp_i (x_{pi}, t_{pi}) \mid \theta_p. \tag{12}$$

Second, it is assumed that the score of a person is a sufficient statistic for their ability:

$$\mathbf{X}_p, \mathbf{T}_p \perp\!\!\!\perp \theta_p \mid \sum_i x_{pi}\left(d_i - t_{pi}\right) - \left(1 - x_{pi}\right)\rho_g, \tag{13}$$

and the total score for an item is a sufficient statistic for the difficulty parameter (δ_i) of an item:

$$\mathbf{X}_i, \mathbf{T}_i \perp\!\!\!\perp \delta_i \mid \sum_p x_{pi}\left(d_i - t_{pi}\right) - \left(1 - x_{pi}\right)\rho_g. \tag{14}$$

Following these assumptions and given the scoring rule,

$$S(x_{pi}, t_{pi}) = x_{pi}\left(d_i - t_{pi}\right) - \left(1 - x_{pi}\right)\rho_g, \tag{15}$$

the joint distribution of response time and response accuracy is of the following form:

$$f(x_{pi}, t_{pi} \mid \theta_p, \delta_i) = \frac{\exp(S(x_{pi}, t_{pi})\vartheta_{pi})}{\int_0^d \sum_{x=1}^0 \exp(S(x_{pi}, t_{pi})\vartheta_{pi})dt}, \tag{16}$$

which derives to the following equation:

$$f(x_{pi}, t_{pi} \mid \theta_p, \delta_i) = \frac{\vartheta_{pi}\exp\Big(\left(x_{pi}\left(d_i - t_{pi}\right) - \left(1 - x_{pi}\right)\rho_g\right)\vartheta_{pi}\Big)}{\exp\left(\vartheta_{pi} d_i\right) - 1 + \vartheta_{pi} d_i \exp\left(-\rho_g \vartheta_{pi}\right)}. \tag{17}$$

Given this joint distribution, the expected score ($\mathbb{E}(S_{pi})$) can be expressed as follows:

$$\mathbb{E}(S_{pi}) = \int_0^d \sum_{x=0}^1 \left(x\left(d_i - t\right) - (1 - x)\rho_g\right) f\left(x, t \mid \theta_p, \delta_i\right) dt \tag{18}$$

Solving this integral results in (6)

Appendix B. Consequences of the FP scoring rule

While this new scoring rule might be more appropriate from an educational point of view, the properties that come with this change are less desirable. For one, the Rasch model is no longer the model for accuracy as is the case for the HSHS rule (Maris & Van der Maas, 2012):

$$p(x_{pi} = 1 \mid \vartheta_{pi}) = \frac{\exp\left(d_i \vartheta_{pi}\right) - 1}{\exp\left(d_i \vartheta_{pi}\right) - 1 + d_i \vartheta_{pi} \exp\left(-p\vartheta_{pi}\right)}. \tag{19}$$

As such, ability is not independent of response accuracy given the number of correct responses. Secondly, the expression for the expected score becomes much more convoluted:

$$\mathbb{E}(S_{pi}) = \frac{-\rho_g \vartheta_{pi}^2 \exp\left(-\rho_g \vartheta_{pi}\right) + \left(\vartheta_{pi} - 1\right)\exp\left(\vartheta_{pi}\right) + 1}{\vartheta_{pi}\left(\exp\left(\vartheta_{pi}\right) - 1 + \vartheta_{pi}\exp\left(-\rho_g \vartheta_{pi}\right)\right)}. \tag{20}$$

Furthermore, while the expected response time for a correct response remains identical to that of the initial HSHS scoring rule:

$$\mathbb{E}(T \mid x = 1) = \frac{1 - \left(d_i \vartheta_{pi} + 1\right)\exp\left(-d_i \vartheta_{pi}\right)}{\vartheta_{pi}\left(1 - \exp\left(-d_i \vartheta_{pi}\right)\right)}, \tag{21}$$

for an incorrect response we find that the expression for the expected response time is only dependent on the item time limit:

$$\mathbb{E}(T \mid x = 0) = \frac{d_i}{2}. \tag{22}$$

The FP scoring rule does not make predictions about response times for incorrect responses given ability levels. In other words, the distribution of expected scores for the FP rule is uniform. As such, the response time for incorrect answers has no influence on ability estimation. If an answer is incorrect, S_{pi} equals the fixed penalty ρ_g. However, because expected response time has no influence on the parameter estimates, the effect on the system performance is negligible.

A final drawback of the switch to the FP scoring rule is that the item selection procedure becomes more complex in comparison to the initial HSHS scoring rule. In the system, items are normally selected with a difficulty such that on average a learner has an expected probability of .75 to give the correct response and the reward for a correct response is equal to the score (S_{pi}). To select a new item, the system calculates the item difficulty value that makes the expected probability of a correct response—based on the learner's ability—equal to a target probability. This target probability is randomly sampled from a normal distribution centered around the chosen difficulty level with a small standard deviation. If we let P denote the target probability correct, for the initial HSHS scoring rule, the expression for the corresponding item difficulty, the target delta δ_i, is given by:

$$\delta_i = \theta_p + \log\left(\frac{1}{P} - 1\right) \tag{23}$$

This is a fairly straightforward expression that is easily computed, after which the system selects the least played item from the 10 items with item difficulties closest to the target delta (δ). For the new FP scoring rule a simple analytical solution does not exist as the target value for the item difficulties will depend on the penalty. Specifically, the target delta in the FP system can be expressed as:

$$\delta_i = \theta_p + \log\left(\frac{1}{P} - 1\right) - \tau_\rho \tag{24}$$

That is, to get the target item difficulty in the new system we need to subtract some penalty dependent constant (τ_ρ) to the expression from Equation (23). The specific values for τ_ρ in the case of an expected probability correct of .75 are illustrated in Fig. 9. From this figure can be seen that in the usual case of $\rho = .1$ we would find that $\tau_\rho \approx 0.546$, which means that we would have to select an item with a lower value of δ in the FP scoring rule compared to the HSHS scoring rule, to get the expected probability correct of .75.

As there exists no analytical solution for obtaining τ_ρ, a practical solution to implement this would be to numerically approximate it for the finite number of penalties used in the system and store these in a database that the system can query and use together with Equation (24) whenever a new item has to be selected.

Data availability

R scripts with analysis are available on GitHub at https://github.com/HankeVermeiren/Psychometrics-of-an-Elo-based-Large-Scale-Online-Learning-System. Data from the Prowise Learn environments are not publicly shared, but access can be requested by contacting the last author.

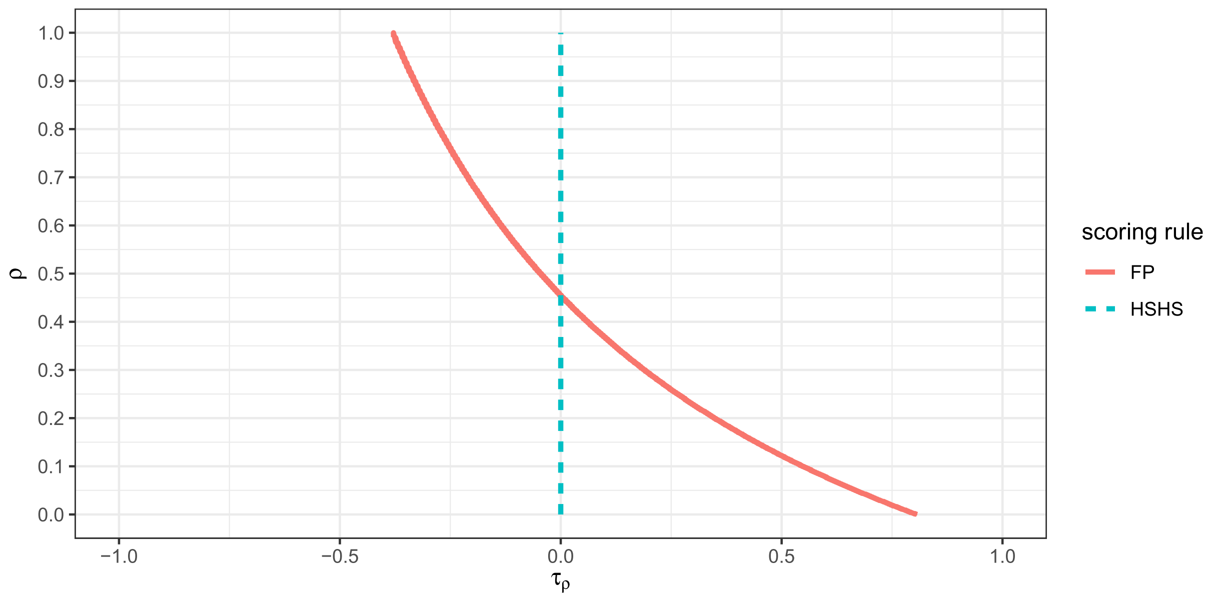

Fig. 9. Visualization of the penalty dependent constant (τ_ρ) required to obtain items for an expected probability correct of .75 with the FP scoring rule.

References

Adedoyin, O. B., & Soykan, E. (2023). Covid-19 pandemic and online learning: The challenges and opportunities. *Interactive Learning Environments*, *31*(2), 863–875.

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)? *Educational Psychology Review*, *33*(4), 1675–1715.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16.

Bolsinova, M., Brinkhuis, M. J., Hofman, A. D., & Maris, G. (2022a). Tracking a multitude of abilities as they develop. *British Journal of Mathematical & Statistical Psychology*, *75*(3), 753–778.

Bolsinova, M., Maris, G., Hofman, A. D., van der Maas, H. L., & Brinkhuis, M. J. (2022b). Urnings: A new method for tracking dynamically changing parameters in paired comparison systems. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *71*(1), 91–118.

Braithwaite, D. W., Goldstone, R. L., van der Maas, H. L., & Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: The case of arithmetic. *Cognition*, *149*, 40–55.

Brinkhuis, M., Cordes, W., & Hofman, A. (2020). Governing games adaptive game selection in the math garden. In *Itm web of conferences, Vol. 33* (p. 03003).

Brinkhuis, M. J., Bakker, M., & Maris, G. (2015). Filtering data for detecting differential development. *Journal of Educational Measurement*, *52*(3), 319–338.

Brinkhuis, M. J., Savi, A. O., 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.

Bureau, J. S., Howard, J. L., Chong, J. X., & Guay, F. (2022). Pathways to student motivation: A meta-analysis of antecedents of autonomous and controlled motivations. *Review of Educational Research*, *92*(1), 46–72.

Chen, J., & Liang, M. (2022). Play hard, study hard? The influence of gamification on students' study engagement. *Frontiers in Psychology*, *13*, Article 994700.

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, *40*(11), 4715–4729.

Coomans, F., Hofman, A., Brinkhuis, M., van der Maas, H. L., & Maris, G. (2016). Distinguishing fast and slow processes in accuracy-response time data. *PLoS ONE*, *11*(5).

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253–278.

Csikszentmihalyi, M. (2013). *Flow: The psychology of happiness*. Random House.

Debeer, D., Vanbecelaere, S., Van Den Noortgate, W., Reynvoet, B., & Depaepe, F. (2021). The effect of adaptivity in digital learning technologies. Modelling learning efficiency using data from an educational game. *British Journal of Educational Technology*, *52*(5), 1881–1897.

de Bree, E., van der Ven, S., & van der Maas, H. (2017). The voice of Holland: Allograph production in written Dutch past tense inflection. *Language Learning and Development*, *13*(3), 215–240.

Deci, E. L., & Ryan, R. M. (2004). *Handbook of self-determination research*. University Rochester Press.

Deci, E. L., & Ryan, R. M. (2013). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.

de Mooij, S. M., Raijmakers, M. E., Dumontheil, I., Kirkham, N. Z., & van der Maas, H. L. (2021). Error detection through mouse movement in an online adaptive learning environment. *Journal of Computer Assisted Learning*, *37*(1), 242–252.

Eggen, T. J., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30*(5), 379–393.

Eglington, L. G., & Pavlik Jr, P. I. (2023). How to optimize student learning using student models that adapt rapidly to individual differences. *International Journal of Artificial Intelligence in Education*, *33*(3), 497–518.

Elo, A. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363.

Eriksson, T., Adawi, T., & Stöhr, C. (2017). "time is the bottleneck": A qualitative study exploring why learners drop out of moocs. *Journal of Computing in Higher Education*, *29*(1), 133–146.

Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *Journal of Defense Modeling and Simulation*, *19*(2), 133–144.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, *48*(3), 377–394.

Gligorea, I., Cioca, M., Oancea, R., Gorski, A.-T., Gorski, H., & Tudorache, P. (2023). Adaptive learning using artificial intelligence in e-learning: A literature review. *Educational Sciences: Theory and Practice*, *13*(12), 1216.

Guay, F. (2022). Applying self-determination theory to education: Regulations types, psychological needs, and autonomy supporting behaviors. *Canadian Journal of School Psychology*, *37*(1), 75–92.

Gütl, C., Rizzardini, R. H., Chang, V., & Morales, M. (2014). Attrition in mooc: Lessons learned from drop-out students. In *Learning technology for education in cloud. Mooc and big data: Third international workshop, ltec 2014, proceedings 3* (pp. 37–48).

Hachey, A. C., Wladis, C., & Conway, K. M. (2023). Investigating online versus face-to-face course dropout: Why do students say they are leaving? *Educational Sciences: Theory and Practice*, *13*(11), 1122.

Heathcote, A., & Matzke, D. (2022). Winner takes all! What are race models, and why and how should psychologists use them? *Current Directions in Psychological Science*, *31*(5), 383–394.

Hilz, A., Guill, K., Roloff, J., Sommerhoff, D., & Aldrup, K. (2023). How to continue? New approaches to investigating the effects of adaptive math learning programs on students' performance, self-concept, and anxiety. *Journal of Intelligence*, *11*(6), 108.

Hofman, A. D., Brinkhuis, M. J., Bolsinova, M., Klaiber, J., Maris, G., & van der Maas, H. L. (2020). Tracking with (un) certainty. *Journal of Intelligence*, *8*(1), 10.

Hofman, A. D., Jansen, B. R., De Mooij, S. M., Stevenson, C. E., & Van der Maas, H. L. (2018a). A solution to the measurement problem in the idiographic approach using computer adaptive practicing. *Journal of Intelligence*, *6*(1), 14.

Hofman, A. D., Kievit, R. A., Stevenson, C., Molenaar, D., Visser, I., & van der Maas, H. L. J. (2018b). *The dynamics of the development of mathematics skills: A comparison of theories of developing intelligence*. OSF Preprints.

Hofman, A. D., Visser, I., Jansen, B. R., Marsman, M., & van der Maas, H. L. (2018c). Fast and slow strategies in multiplication. *Learning and Individual Differences*, *68*, 30–40.

Jansen, B. R., Hofman, A. D., Savi, A., Visser, I., & van der Maas, H. L. (2016). Self-adapting the success rate when practicing math. *Learning and Individual Differences*, *51*, 1–10.

Jansen, B. R., Louwerse, J., Straatemeier, M., Van der Ven, S. H., Klinkenberg, S., & Van der Maas, H. L. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences*, *24*, 190–197.

Jeno, L. M., Nylehn, J., Hole, T. N., Raaheim, A., Velle, G., & Vandvik, V. (2023). Motivational determinants of students' academic functioning: The role of autonomy-support, autonomous motivation, and perceived competence. *Scandinavian Journal of Educational Research*, *67*(2), 194–211.

Kandemir, E. N., Vie, J.-J., Sanchez-Ayte, A., Palombi, O., & Ramus, F. (2024). Adaptation of the multi-concept multivariate elo rating system to medical students' training data. In *Proceedings of the 14th learning analytics and knowledge conference* (pp. 123–133).

Karsen, M., Rahman, S. B. A., Kurniawan, Y., et al. (2023). Incompletion rate factors in mooc: A systematic literature review. In *2023 international conference on university teaching and learning (incult)* (pp. 1–6).

Kaya, O. S., & Ercag, E. (2023). The impact of applying challenge-based gamification program on students' learning outcomes: Academic achievement, motivation and flow. *Education and Information Technologies*, *28*(8), 10053–10078.

Kew, S. N., & Tasir, Z. (2022). Learning analytics in online learning environment: A systematic review on the focuses and the types of student-related analytics data. *Technology, Knowledge and Learning*, *27*(2), 405–427.

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In *Proceedings of the 2009*. gmac conference on computerized adaptive testing (Vol. 2).

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education*, *57*(2), 1813–1824.

Lee, Y., & Choi, J. (2011). A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, *59*, 593–618.

Li, K. C., & Wong, B. T.-M. (2023). Features and trends of personalised learning: A review of journal publications from 2001 to 2018. *Personalized Learning*, 4–17.

Mangaroska, K., Vesin, B., & Giannakos, M. (2019). Elo-rating method: Towards adaptive assessment in e-learning. In *2019 ieee 19th international conference on advanced learning technologies (icalt), Vol. 2161* (pp. 380–382).

Maris, G., & Van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633.

Martin, F., Chen, Y., Moore, R. L., & Westine, C. D. (2020). Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development*, *68*, 1903–1929.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, *23*(3), 187–194.

Mercer, M., & Gulseren, D. B. (2024). When negative feedback harms: A systematic review of the unintended consequences of negative feedback on psychological, attitudinal, and behavioral responses. *Studies in Higher Education*, *49*(4), 654–669.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74.

Mulder, E., van de Ven, M., Segers, E., Krepel, A., de Bree, E. H., van der Maas, H., de Jong, P. F., & Verhoeven, L. (2021). Serious game-based word-to-text integration intervention effects in English as a second language. *Contemporary Educational Psychology*, *65*, Article 101972.

Ooge, J., De Braekeleer, J., & Verbert, K. (2024). Nudging adolescents towards recommended maths exercises with gameful rewards. In *International conference on artificial intelligence in education* (pp. 328–335).

Ou, L., Hofman, A. D., Simmering, V. R., Bechger, T., Maris, G., & van der Maas, H. L. (2019). Modeling person-specific development of math skills in continuous time: New evidence for mutualism. *International Educational Data Mining Society*.

Papousek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. In *Educational data mining 2014*.

Park, J. Y., Cornillie, F., Van der Maas, H. L., & Van Den Noortgate, W. (2019a). A multidimensional irt approach for dynamically monitoring ability growth in computerized practice environments. *Frontiers in Psychology*, *10*, 620.

Park, J. Y., Joo, S.-H., Cornillie, F., van der Maas, H. L., & Van den Noortgate, W. (2019b). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behavior Research Methods*, *51*, 895–909.

Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis–a new alternative to knowledge tracing. In *Proceedings of the 14th international conference on artificial intelligence in education*.

Pelánek, R. (2016). Applications of the elo rating system in adaptive educational systems. *Computers and Education*, *98*, 169–179.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction*, *27*, 313–350.

Pressey, S. L. (1927). A machine for automatic teaching of drill material. *School & Society*.

Ranger, J., & Kuhn, J.-T. (2012). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement*, *36*(3), 214–231.

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*, 491–513.

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

Schaeffer, C. E., & Konetes, G. D. (2010). Impact of learner engagement on attrition rates and student success in online learning. *International Journal of Instructional Technology and Distance Learning*, *7*(5), 3–9.

Shaikh, U. U., & Asif, Z. (2022). Persistence and dropout in higher online education: Review and categorization of factors. *Frontiers in Psychology*, *13*, Article 902070.

Sinharay, S., Haberman, S. J., & Lee, Y.-H. (2011). When does scale anchoring work? A case study. *Journal of Educational Measurement*, *48*(1), 61–80.

Skinner, B. F. (1958). Teaching machines: From the experimental study of learning come devices which arrange optimal conditions for self-instruction. *Science*, *128*(3330), 969–977.

Straatemeier, M. (2014). *Math garden: A new educational and scientific instrument*. Doctoral Dissertation. Universiteit van Amsterdam.

ten Broeke, N., Hofman, A. D., Kruis, J., de Mooij, S., & van der Maas, H. L. J. (2021). *Predicting and reducing quitting in online learning*. OSF Preprints.

Van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Van der Linden, W. J., & Hambleton, R. K. (2016). *Handbook of item response theory*. Chapman and Hall/CRC.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842.

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

van der Ven, S. H., van der Maas, H. L., Straatemeier, M., & Jansen, B. R. (2013). Visuospatial working memory and mathematical ability at different ages throughout primary school. *Learning and Individual Differences*, *27*, 182–192.

Vandierendonck, A. (2021). On the utility of integrated speed-accuracy measures when speed-accuracy trade-off is present. *Journal of Cognition*, *4*(1).

Van Duijvenvoorde, A. C., Zanolie, K., Rombouts, S. A., Raijmakers, M. E., & Crone, E. A. (2008). Evaluating the negative or valuing the positive? Neural mechanisms supporting feedback-based learning across development. *The Journal of Neuroscience*, *28*(38), 9495–9503.

Vermeiren, H., Bolsinova, M., van der Maas, H., & Van Den Noortgate, W. Balancing stability and flexibility: Investigating a dynamic k value approach for the elo rating system in adaptive learning environments. Preprint, https://doi.org/10.31234/osf.io/6hzke.

Vermeiren, H., Hofman, A., Bolsinova, M., Van Den Noortgate, W., & van der Maas, H. Computerized adaptive literacy learning. Preprint, https://doi.org/10.31234/osf.io/5ubth.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2023a). When adaptive learning is effective learning: Comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments*, *31*(2), 793–803.

Wang, W., Zhao, Y., Wu, Y. J., & Goh, M. (2023b). Factors of dropout from moocs: A bibliometric review. *Library Hi Tech*, *41*(2), 432–453.

Weller, M. (2018). Twenty years of edtech. *Educause Review Online*, *53*(4), 34–48.

Zeybek, N., & Saygı, E. (2024). Gamification in education: Why, where, when, and how?—a systematic review. *Games and Culture*, *19*(2), 237–264.

Zhang, B., Shi, Y., Li, Y., Chai, C., & Hou, L. (2023). An enhanced elo-based student model for polychotomously scored items in adaptive educational system. *Interactive Learning Environments*, *31*(9), 5477–5494.

Zhang, K., & Aslan, A. B. (2021). Ai technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, *2*, Article 100025.