Using Empirical Learning Curve Analysis to Inform Design in an Educational Game

Erik Harpstead and Vincent Aleven
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA, USA 15232
{eharpste, aleven}@cs.cmu.edu

ABSTRACT

Having insights into players' learning has important implications for design in an educational game. Empirical learning curve analysis is an approach from intelligent tutoring systems literature for measuring student learning within a system in terms of the skills involved. The approach can be used to evaluate how well different hypothesized models of required skills fit to actual student performance data from the game. This information can be used to highlight whether players need more practice with specific concepts, how the game's progression might be altered, and whether the game is succeeding at its educational objectives. In this paper we apply empirical learning curve analysis to *Beanstalk,* an educational game designed to teach young children the concept of balance. We show that the process is able to give insight into the detailed skills and concepts (or knowledge components) that players are learning, and give implications for level (re)design by highlighting a previously unforeseen shortcut strategy.

Author Keywords

Educational Games; Learning Curves; Learning

ACM Classification Keywords

H.5.2 User Interfaces: Evaluation/methodology; K.3.1 Computer Uses in Education: Computer-assisted instruction (CAI); K.8.0 Personal Computing: Games.

INTRODUCTION

The link between play and learning has been cited by many researchers [17,24,31,36]. In recent years, there has been a renewed focus on the design and development of educational games that foster learning through play [13,27]. This has led researchers to investigate the dynamics of learning within games and how learning can best be measured [8,25,26].

Measurement of learning in educational games is often considered an end in itself, as a way to prove that the game actually fosters learning. In these cases it is common to measure a change in players' aptitude with external pre-posttests [5]. This orientation is based on the desire to measure skill transfer outside of the game to real world tasks, something often assumed but not guaranteed by a game's design [29].

Learning measurement need not be focused purely on the summative context. When learning is measured throughout the game development process it can provide useful insight to designers [25]. Understanding how players are progressing in learning the skills targeted by a game can have many design implications. For example, it can highlight which design variations lead to more efficient learning [28].

Game user research has explored the dynamics of skill learning over time. By looking at the number of mechanics in each level and the rate at which new mechanics are introduced a rough learning curve can be established for a game [26]. Other work has also approached the issue of learning dynamics by examining the patterns of breakdowns and breakthroughs that players experience as they learn a new game [20].

The field of intelligent tutoring systems research has developed many methods for helping educational technology researchers understand the dynamics of learning within an educational environment. While there has been some application of these methods to games [9,28,32], little prior research has explored how measurement and modeling of learning can inform the design process. Particularly, none of these prior applications have looked at how the modeling process can help designers see where they might be wrong about the skills they believe to be targeted by their game and how altering the designer's skill model can highlight previously unforeseen skills, a process referred to as model refinement [41].

In this paper we present an application of established educational data mining methods, which are normally applied to log data from intelligent tutoring systems, to an educational game. This process takes player performance data (from game play logs) and uses a combination of statistical modeling and visualization to help designers understand what skills are exercised by players in their game. The approach lends itself to actionable insight by

CHI PLAY 2015, October 03 - 07, 2015, London, United Kingdom

ACM 978-1-4503-3466-2/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2793107.2793128

highlighting skills that are under practiced and by exposing skills that designers may not have expected to be present. We demonstrate this approach in the context of the educational game *Beanstalk* [3,12] and discuss how the approach might be adapted to other game contexts, including ones that are not explicitly educational.

EMPIRICAL LEARNING CURVE ANALYSIS

Empirical Learning Curve Analysis is a method for evaluating interactive educational software from the intelligent tutoring systems tradition [4]. The method is based on classical cognitive theory by Newell and Rosenbloom [34]. According to their theory, as people are given more opportunities to practice the use of a particular skill or concept their chance of incorrectly applying the skill or concept should decrease according to a power law. While there has been some debate whether a power law or an exponential law is more appropriate [18], the general notion that error rate decreases over opportunities in a non-linear fashion holds.

In intelligent tutoring literature, empirical learning curve analysis is facilitated by formalizing the skills and concepts required for a task into a Knowledge Component (KC) model [4]. A KC represents a specific unit of cognitive function, such as a procedural skill or element of factual knowledge, that is necessary to successfully perform a given task [23]. KC models are created by assigning labels of the skills or concepts required for each step of a task. This process is often done through a combination of empirical (e.g. having experts think aloud while performing the task) and theoretical (e.g. rationally constructing a set labels) task analysis [4]. This process is similar to the practice of enumerating the skills required on different levels of a game, which is advocated by some game designers [36].

Empirical Learning Curves are traditionally plotted with the number of opportunities to practice a KC along the x-axis and the average error rate of all learners on the y-axis. The intercept of the curve represents the initial difficulty of a KC, for the given student population, while the slope of the curve represents the rate at which learners appear to be mastering the KC. If the curve is steep then learners are reaching mastery quickly, whereas if the curve is shallow or flat, learning is happening slowly, or possibly not at all.

DataShop is a data repository and tool suite that is commonly used to apply empirical learning curve analysis to instructional technologies [22]. DataShop provides a number of useful tools for visualizing and interacting with learning curves based on log data from users in terms of different KC models. It also provides a workflow for statistically fitting a given KC model to student data. This provides insights into the difficulty and learning rates of different skills under a particular KC model and makes it possible to formally compare which model is a better way of explaining the skills involved in a task.

The Additive Factors Model (AFM) [10] is a statistical model commonly used to evaluate different KC models in terms of real learner data. AFM is a specialized form of logistic regression that uses student success at a step as a 0 or 1 dependent variable. The regression uses three independent variable terms: (1) student intercepts, modeled as a random effect, which captures the assumption that all students come in with differing amounts of prior knowledge; (2) KC intercepts, which account for the assumption that different KCs will have different initial difficulty; and (3) KC slope, modeled as an interaction effect between the KC and number of practice opportunities, which encodes the assumption that all learners will generally increase in ability with a KC at a similar rate given the same number of opportunities. The KC slope term is constrained to be greater than or equal to 0, meaning that students cannot "unlearn" a KC from experience. The regression equation takes the following form:

$$\ln\frac{p_{ij}}{1-p_{ij}} = \theta_i + \sum_k \beta_k Q_{kj} + \sum_k Q_{kj}(\gamma_k N_{ik})$$

In this equation, p_{ij} is the probability that student i gets step j correct. θ_i represents the competency (or intercept) of student i. Q represents a binary matrix mapping KC k to step j. β_k represents the intercept parameter for KC k (i.e., the ease of this KC). γ_k represents the slope (i.e., learning rate) for KC k and N_{ik} represents the number of opportunities student i has had to practice KC k .

AFM can be seen as a generalization of the Item Response Theory Rasch Model [35] but accounting for learning. If the KC by opportunity interaction (i.e., $\sum_k Q_{kj}(\gamma_k N_{ik})$) is left off and the KC model is defined with every unique step being its own KC then it is equivalent to the Rasch Model.

BEANSTALK

The game that we will be discussing is called *Beanstalk* [3,12]. *Beanstalk* (Figure 1) is a game designed to teach the basic physical properties of a balance beam to young children (5-8 year olds). The game is informed by classical cognitive theory by Siegler where it was found that children have trouble learning to integrate the properties of weight and distance from the fulcrum when judging what will happen to a simple balance beam system [39].

In the game players must help Jack (or Jackie) return a teddy bear to the monster that lives in the sky by keeping the beanstalk balanced while it grows. Occasionally, bugs will fall onto the beam at the top of the beanstalk, causing it to fall out of balance and impede Jack's progress to the sky. The player must add weight to counteract the bugs by growing flowers (bugs and flowers are assumed to have the same mass). Players are constrained by having a limited amount of water with which to grow flowers and also by having some flower positions unavailable for growing, depending on the specific level design. Once the player has run out of water the beam comes to rest according to the

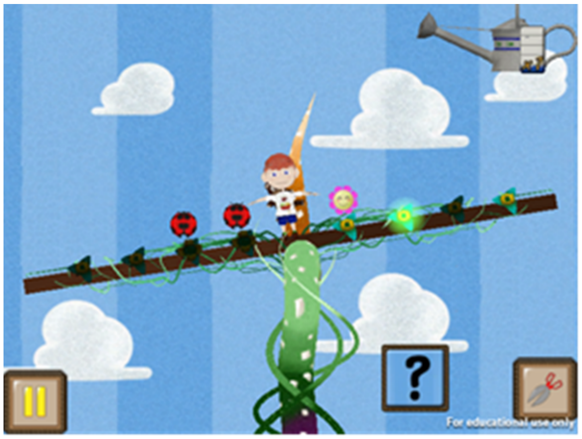

Figure 1. A screenshot of *Beanstalk*. The player is choosing to grow a second flower on the right to try and counteract the weight of the two red bugs on the left.

weights on either side. If the beam is balanced then the player proceeds on to the next level; otherwise, the player loses and must retry the same level.

The game levels are organized into a series of 7 tiers and a simple mastery learning paradigm is used to control advancement. Players must complete at least 9 levels per tier and successfully solve a number of levels in a row before advancing to the next tier. If a player has not satisfied both advancement conditions, then they proceed to an extra set of levels until they can achieve the winning streak criteria.

According to prior research by Siegler [39] children progress through four stages of development on the way to mastering the balance beam task. Firstly, children primarily fixate on how much weight is on either side of the beam, regardless of its position. Second, children will begin to take into account each weight's distance from the fulcrum if weight is the same. Third, children begin to realize that there are two factors at play, but do not yet know the rule that balance is governed by a sum of cross products of weight and distance. This misconception will bias their thinking toward whichever feature is more exaggerated. When neither weight nor distance is clearly exaggerated, stage three children will generally guess; such examples are referred to as *conflict cases*. Finally, at the fourth stage, children have learned that the balance beam is governed by the sum of cross products between distance and weight on both sides. *Beanstalk*'s goal is to advance players through these stages of development, ideally ending in stage four.

The *Beanstalk* level designers were faced with the challenge of coming up with a sequence of levels that support Siegler's developmental sequence. Although it is helpful to have a theory that states how children naturally progress in their learning for a domain such a theory does not describe how to create instruction, that is, it does not specify what learning experiences (i.e., level sequence) would be most helpful in helping learners efficiently acquire robust knowledge. Further, even with a theory, the effect of carefully designed game levels on students' learning cannot be fully understood in advance.

ANALYSIS

Our analysis of *Beanstalk* focuses on the question: What KC model best accounts for the data? Answering this question enables us to understand student learning with the game in substantial detail, which in turn helps us understand whether the design of the game is successful in accomplishing its educational goals. The analysis follows in two stages. Firstly, we analyze learning with reference to an initial set of baseline KC models, which include naïve models as well as coarse-grained cognitively plausible models. Next, we look at different refinements of one of the latter models as a way of exploring hypotheses about difficulty and learning. These model variations come from rational analysis of the task at hand, and through exploration of the data. As discussed below, we compare models in terms of their fit with the game play data (whether students solved each level correctly) and their accuracy in predicting unseen data.

To analyze *Beanstalk* we uploaded a sample of player log data to DataShop. This data was captured as part of a formative evaluation of the game. The evaluation was performed in classrooms from multiple Pittsburgh area public school with children in the target demographic (ages 5-8). 177 students were given two 20-minute periods to play the game in class, with an average of 35.5 total minutes played per student.

The original log data contains data from 12,007 level attempts by players; however, some data was removed because of systematic bias (e.g., levels that – by design – are 100% successful because they are designed as "worked examples" that introduce a new complexity to players, e.g. the first conflict case level is a worked example. To accomplish their purpose these levels are designed to be impossible to fail. While these levels do involve particular KCs we remove them from our analysis to prevent this bias from affecting results. This removal results in 10,330 level attempts in the entire sample.

Base Modeling

We begin the KC modeling process by developing a set of base models. These models serve as the backbone to exploring different (more fine-grained) models of the skills involved in playing *Beanstalk*. To start, we created two naïve KC models, one coarse-grained, and one more fine-grained, that represent "boundaries" on the different ways that a game's designers might think of skill within a game, but are not thought of as cognitively plausible or desirable.

The first model is called the Single-KC (1 KC) model. This model assumes that there is only one knowledge component present in the entire dataset; this could be interpreted as "skill at balancing a beam" or alternatively "skill at playing *Beanstalk*". While it is colloquial to say that a person has

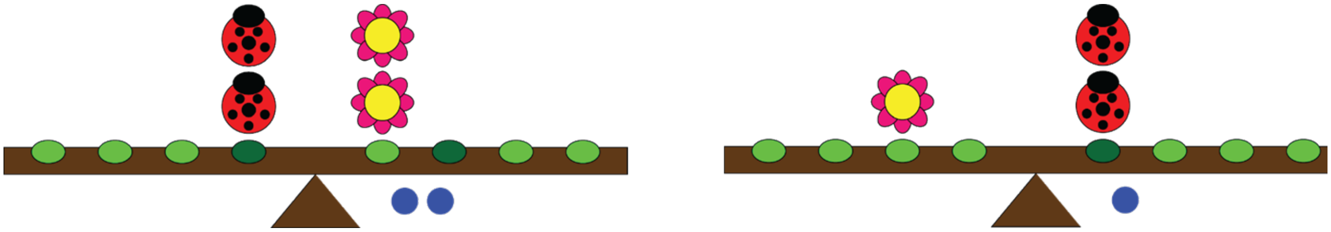

Figure 2. Level specifications for a mirror (M) level (left) and conflict-balance (CB) level (right). The bugs are given as the initial conditions of the level, the flowers represent the player's solution, and the blue circles below the beam represent the amount of water initially available to the player.

aptitude at an entire game, e.g. saying someone is good at chess, it is almost never the case that a task involves a single monolithic skill [23]. It is far more likely that different sub-tasks invoke different kinds of skills or knowledge, e.g. in chess deciding which piece to move requires different skills than recalling the legal forms of movement for a given piece. Cognitive theories like ACT-R [7] and KLI [23] postulate that knowledge or skills within a task domain are built out of small knowledge components, which each need to be learned separately through practice.

The second naïve model we use is called the Naïve-Tier (7 KC) model. This model takes from the designers' specification that each of the 7 tiers of the game was designed around a common theme, e.g. tier 1 is mostly simple levels meant to introduce the game mechanics, while more complex balance configurations fit into later tiers. It then assigns a KC label to each level according to the tier it is in. Generally, each tier is intended to be harder than the one preceding it, and each tier introduces new concepts. This would be analogous to assuming that every new tier of a game like *Angry Birds* exercises a unique skill. While it may be true that a new section of a game adds mechanics that require new skills to succeed, it is rare that no previous skills are required or that a single skill is used across all the levels of a tier.

For the final base model we turn more directly to the literature from Siegler. In his original study Siegler had a taxonomy of balance beam judgment items informed by the stages learners go through in understanding the balance beam [39]. In developing our model we look at each level of the game and consider both what is present to the player initially and what the designers considered to be the solution to the level. We then take this solution state and consider it in terms of Siegler's taxonomy. There are multiple ways to solve many levels, but to demonstrate the utility of KC modeling to assess a designer's intuitions against player data we use the designer envisioned solutions in model creation. This is akin to a rational task analysis for generating KC models [4].

Model Name	KCs	AIC	BIC	CV-RMSE
Single-KC	1	14276.19	15574.36	.4950
Naïve-Tier	7	13118.61	14503.81	.4780
Siegler	2	*12119.20*	*13430.15*	*.4479*

Table 1. Fit statistics for the 3 base models. Cross-Validated Root Mean Squared Errors (CV-RMSE) is calculated using 3-fold cross-validation.

From Siegler's original study the main distinction between balance beam tasks is balance versus conflict-balance. A balance level, which we refer to as a mirror (M) level, is one where the positions and counts of weights is equal on both sides of the beam when using the designer envisioned solution. A conflict-balance (CB) level is one where the positions or counts of weights are different on each side of the beam when using the designer envisioned solution (see Figure 2 for an illustration of these concepts). Under this Siegler (2 KC) model, we would expect that mirroring levels have a lower initial difficulty and are easier to master while CB levels are initially harder and slower to master.

Base Model Evaluation

Once each KC model has been created and player log files have been tagged with their corresponding labels we upload the data to DataShop and use AFM (described above) to evaluate the models according to their fit to player data or their predictive accuracy on held-out data. Throughout this paper we report the fit of KC models using Akaike Information Criterion (AIC) [1], Bayesian Information Criterion (BIC) [37], and cross-validated root mean squared error (CV-RMSE) stratified by level. AIC and BIC are both standard measures of model fit that penalize models for having larger numbers of parameters; in the case of KC modeling, this is the number of KCs in the model. Both metrics have an arbitrary scale with strictly lower values representing a better fitting model. CV-RMSE partitions the data into three folds ensuring no level's data is split between folds. Once the data is partitioned the model is trained on two folds and then used to predict the values of the remaining fold. The root mean square of the error between the prediction and actual values is then reported, with smaller values indicating a more accurate model.

When evaluating KC models the different fit statistics often agree with each other, but in cases when they do not it is useful to understand how their conclusions differ. AIC and BIC more strongly punish for having too many KCs, with BIC being stronger in this regard. The CV-RMSE values are arguably more rigorous measurements of fit because they evaluate predictive ability. Level-stratified CV-RMSE

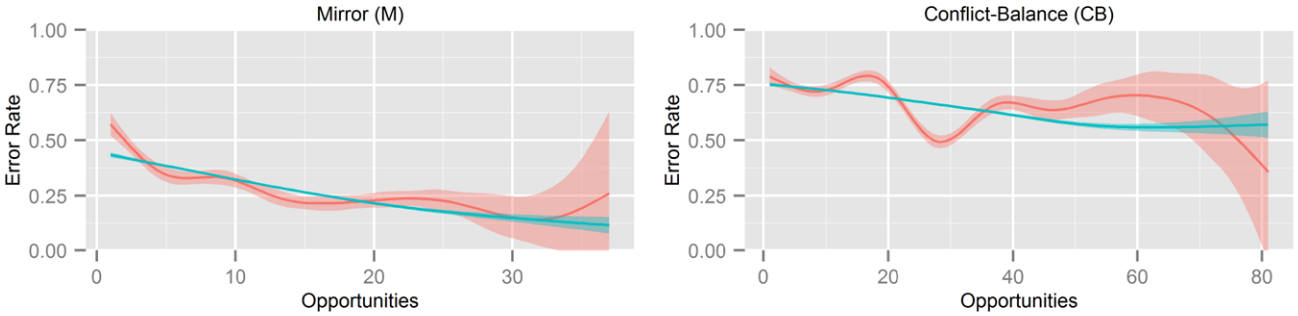

Figure 3. Learning curves for the Mirror (M) and Conflict-Balance (CB) KCs of the Siegler based KC model. The red line plots the actual player error rate at each opportunity while the blue line plots the curve fit by AFM. The shaded regions on both lines denote the standard errors of the data.

is more useful for evaluating the model's ability to predict new items. Additionally, item stratification gives a sense of how well a KC transfers between items within a tutor, or levels in a game.

The fit statistic results for the base models can be seen in Table 1. From the results, the Siegler model is the best fitting model across all metrics. This suggests that the Siegler based KC model is the most appropriate description of the skills involved in *Beanstalk* when compared to the two naïve models.

Taking the model parameter results of the Siegler based model fit by AFM, we can plot the learning curves shown in Figure 3. Throughout this paper all learning curves are rendered using a loess smoothing [14] over both the actual player error rate and the error rate predicted by AFM. All curves are also rendered with standard error bounds. This is done to denote the drop off in the player population as the opportunity count increases – this drop off occurs because players who master a skill will have advanced on to new problems. This drop off explains the much greater standard error on later opportunities clearly shown in Figure 3.

In Figure 3, each curve shows a decrease in error rate on M and CB levels as players are given opportunities to practice each skill, i.e. attempt a level for that KC. The patterns we see fit with our hypothesized dynamics where M levels have a lower initial difficulty (KC intercept ~44% error) and are easier to master (KC slope .056). In contrast the CB levels are initially more difficult (KC intercept ~77% error) and are harder to master (KC slope .008).

While these results generally concur with those of Siegler they have some implications in terms of *Beanstalk*'s design. For example, while the CB KC does appear to demonstrate a pattern of learning it does not converge to a very promising value, roughly 55.76% error by most players' last opportunity to practice. From these results it could be concluded that the game seems to do well with mirroring cases but does not help students master the conflict-balance concept, perhaps because it does not provide enough practice. However, this assumes that the 2KC Siegler model is the best explanation of skill in *Beanstalk*. It is hard to tell what design recommendations could follow from this conclusion because the CB KC labels roughly half of the game's levels. A finer grained analysis can be used to provide more actionable insight.

Model Variation

To better understand what can be done with the conclusions of the 2KC Siegler model we explore how variations of the model impact our measured patterns of learning in the game. As a first step we explore creating a finer grained model by splitting the Siegler model to better understand what is causing the high level learning patterns in the M and CB KCs. To do this we apply a rational task analysis approach to *Beanstalk* to create a cognitively informed variation of the model. Next, we further investigate the differences between the models' predicted values and the actual error rate to find whether encoding levels with KCs that capture a strategy no foreseen by the designers might better account for the data.

Base Model Elaboration

In this first variation we test whether a more fine-grained knowledge model might reflect players' psychological reality better. To this end we examined an elaboration of Siegler's original taxonomy by looking at the number of positions involved on each side of the beam. Because balance is ultimately governed by a sum of cross products rule we posit that levels that require the integration of multiple peg positions per side are more cognitively demanding than ones that do not. Under this model a particular balance level can have either single or multiple pegs being used on either the given side (i.e. the side that starts with bugs on it) or the acting side (i.e. the side that the player works on). The resulting model has 6 KCs:

- **M-GSP-ASP** – mirroring with a single peg used on each side of the beam (used on 12 levels).
- **M-GMP-AMP** – mirroring with multiple pegs used on each side of the beam (used on 21 levels).
- **CB-GSP-ASP** – conflict-balance with a single peg used on each side of the beam (used on 19 levels).

- **CB-GSP-AMP** – conflict-balance with a single peg on the given side of the beam and multiple pegs on the acting side of the beam (used on 9 levels).
- **CB-GMP-ASP** – conflict-balance with multiple pegs on the given side of the beam and a single peg on the acting side (used on 6 levels).
- **CB-GMP-AMP** – conflict-balance with multiple pegs used on both sides of the beam (used on 14 levels).

Under this Siegler+Pegs (6 KC) model we would hypothesize that levels with single pegs have a lower initial difficulty and are easier to learn than ones involving multiple pegs. This is because the mental calculation to solve the balance requires an extra addition step when there are multiple pegs on a side, whereas a single peg side only requires a single multiplication step.

When fit with AFM this finer grained model outperforms the original Seigler model on all statistics (see Table 2). Looking at the plotted curves (Figure 4) we find that our original hypothesis is not upheld by this model. Levels involving multiple pegs (MP) appear to have lower initial difficulties than those involving single pegs (SP), particularly when the multiple pegs appear on the acting side of the beam (AMP). One way to explain the difference in initial difficulty is that single peg levels generally precede levels with multiple pegs. The game begins with single peg levels and the first time conflict cases are introduced is with single peg levels, suggesting there is some amount of balance dynamics and game mechanics learning accounted for by the single peg KCs. However, the two curves which seem to most strongly support the conclusion that multiple pegs is harder (CB-GSP-AMP, and CB-GMP-AMP) do not appear to demonstrate a strong fit between the predicted and actual values.

Model Name	KCs	AIC	BIC	CV-RMSE
Siegler	2	12119.20	13430.15	.4479
Siegler+Pegs	6	11845.57	13214.46	.4501
SP+Rote	7	*11050.39*	*12433.76*	*.4205*

Table 2. Fit statistics for the two variants and the original Siegler model. Cross-Validated Root Mean Squared Error (CV-RMSE) calculated using 3-fold cross-validation.

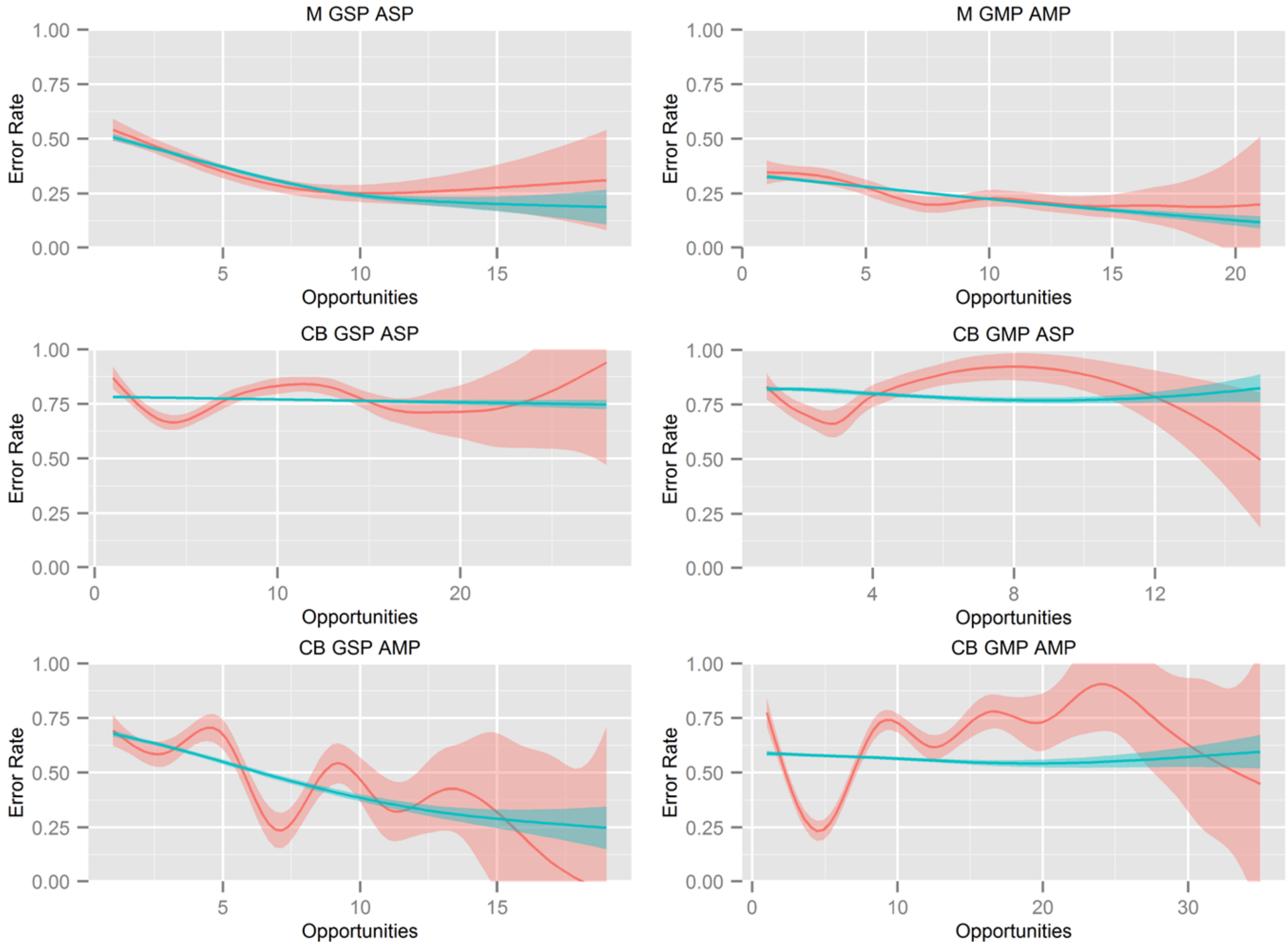

Figure 4. Learning Curves for the Siegler+Pegs model. The red line plots the actual player error rate at each opportunity while the blue line plots the curve fit by AFM. The shaded regions on both lines denote the standard errors of the data.

Another pattern visible from the learning curves that parallels the original Siegler model is that learning appears to be happening in mirroring (M) KCs but not in conflict-balance (CB) KCs, with the exception of CB-GSP-AMP. This would mean that players tend to be learning the pattern in CB-GSP-AMP levels but not in the other conflict levels. This could be because players are simply not given enough practice with conflict-balance items, suggesting that the design team might need to design more levels. Alternatively, this pattern could be from a more nuanced aspect of the design not communicating balance concepts correctly. This pattern warrants a deeper exploration of the *Beanstalk* data.

Alternative Strategy Discovery

A notable pattern visible in the learning curves fit by the Siegler+Pegs model is that the CB-GSP-AMP and CB-GMP-AMP curves do not fit the actual performance data well. This is evidenced by the particularly jagged appearance of the actual player data (solid red line) relative to the fit model's curve (dashed blue line). In empirical learning curve analysis this kind of pattern is characteristic of a hidden skill or difficulty factor [6]. This suggests that while our current model considers a set of levels to exercise a particular KC it is possible that a subset of those levels actually exercises a different KC that we have not yet considered. The Conflict-Balance curve of the original Siegler model (Figure 3) has a similar character.

To get a better understanding of what knowledge components the CB-GSP-AMP and CB-GMP-AMP levels involve, we can look at model fit on a level-by-level basis rather than a KC basis, as we have done so far. This is done by examining the levels that have a large residual error rate (i.e. difference between the model's prediction and the actual error rate observed in players). DataShop provides a tool for this kind of analysis called the Performance Profiler. Figure 5 shows a performance profiler view of the *Beanstalk* levels with the six highest and six lowest residual error rates. Each bar shows the actual error rate on a level while the blue line shows the predicted error rate. The levels to the left of the figure are substantially easier than our current skill model would predict, while the levels to the right of the figure are harder than expected. This pattern suggests that these levels particular would benefit from a more thorough investigation to see if some aspect of their design exercises a different skill than we may have thought.

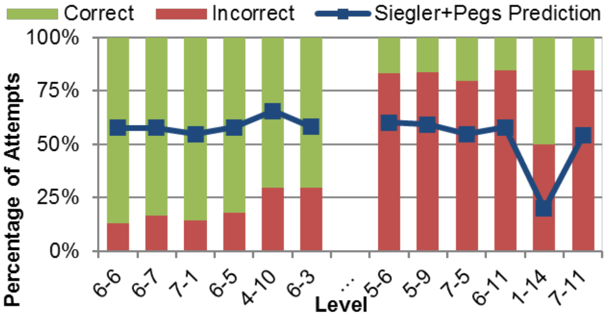

Figure 6. A plot of the actual and predicted error rates for the top and bottom 6 levels sorted by residual.

When examining the level design specifications for the levels that are substantially easier than expected, we found that all of the levels could be solved using a simple strategy of putting a single weight on every open peg (see Figure 6 for examples). This is a form of shallow strategy in that it sometimes leads to correct answers to *Beanstalk* levels, but it is not hard to construct cases in which it would lead players astray (in fact, *Beanstalk* has some levels in which the rote strategy leads to wrong answers). Using the rote strategy does not reflect an understanding of the underlying physics principle. Avoiding this kind of shallow learning and promoting acquisition of appropriately contextualized knowledge is a key goal in education [2].

We developed a simple script to parse the game's level specifications and found all levels that could be solved using the rote strategy. We created the SP+Rote (7 KC) model that re-categorizes levels where the simple procedure could be applied with a "Rote Strategy" KC. This relabeling affected 4 M-GMP-AMP, 2 CB-GSP-ASP, 1 CB-GMP-ASP, 3 CB-GSP-AMP, and 5 CB-GMP-AMP levels. We then re-ran AFM using this new KC model.

This new model fits the player data better than the original Siegler model and the Siegler+Pegs model (see Table 2). This result would suggest that it is more accurate to consider these levels as exercising the rote solution strategy rather than the domain relevant KC they were previously assumed to exercise. This change in KC label also has an effect on the curves fit by the model (see Figure 7). The rote solution strategy primarily appears on levels in the CB-GSP-AMP and CB-GMP-AMP categories. When the rote strategy is separated out as its own KC is has the effect of smoothing out the two previously jagged curves, adding credence to the interpretation that this model is a more reasonable explanation of skill in the game.

The change in the CB learning curves also has an effect on our interpretation of the model with regard to our hypothesis of more pegs meaning more difficulty. When the Rote Strategy KC is present in the model all of the CB related curves shift to having roughly equal intercepts

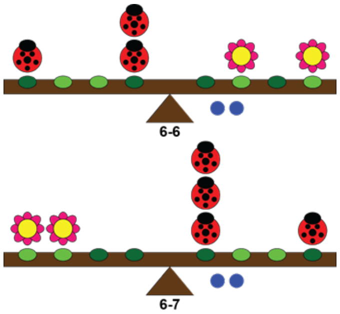

Figure 5. The specification for two levels that can be solved using the rote solution strategy.

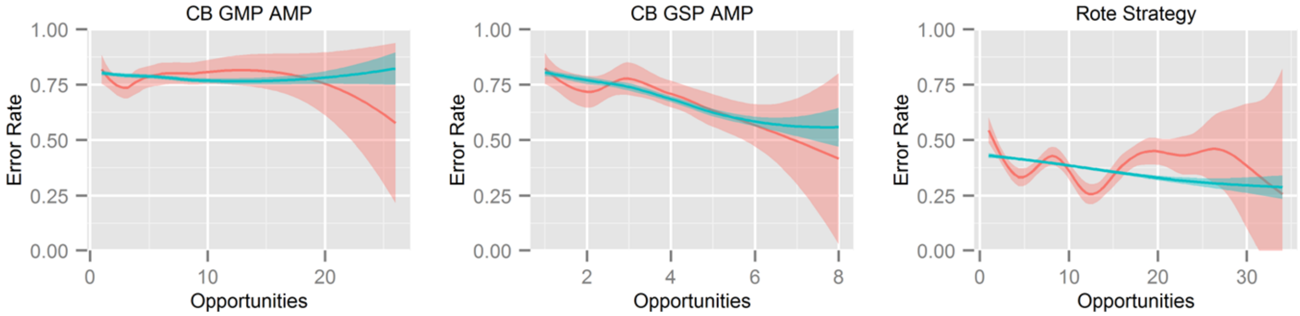

Figure 7. Learning curves altered by adding the Rote Strategy KC to the Siegler+Pegs model (other curves remain unchanged). The red line plots the actual player error rate at each opportunity while the blue line plots the curve fit by AFM. The shaded regions on both lines denote the standard errors of the data.

(~84% error rate on average). This would imply that, within the context of the rote solution strategy, knowing about the number of pegs involved on either side of a beam does not provide insight into the initial difficulty of a conflict-balance level.

One pattern in the data that is still not explained is the fact that players appear to be learning the CB-GSP-AMP KC but none of the other conflict-balance KCs. When examining the levels that fit the CB-GSP-AMP category, after accounting for the rote solution strategy, a simple pattern does become apparent from their configurations. In all of these levels, players do not have to stack flowers, i.e. place more than one weight on a peg, to solve the balance. This has a simplifying effect on the sum of cross products calculation that governs the balance beam by turning one side of the equation into pure addition. Contrary to what might have been expected from a purely rational cognitive perspective, these levels are likely easier for players to learn because of this simplification. The game designers could benefit from this information and move these levels earlier to ease the introduction of the conflict-balance concept.

Overall, the patterns present from the rote solutions strategy would suggest that the designers might do well to consider removing or altering levels that can be solved with the strategy. One approach might be to try to design the levels in such a way that this strategy can never succeed. Presumably this would lead students to not develop the strategy in the first place. An alternative or supplementary approach may be to extend the game so that it better supports understanding of the balance formula. The prevalence of the rote strategy (which reflects a lack of understanding) suggests that learning the sum of cross products formula purely inductively may be very difficult for the target audience and may require further scaffolding.

DISCUSSION

Our analysis throughout this paper has facilitated a clearer understanding of the dynamics of learning in the game *Beanstalk*. Initially, we found that a model informed by the original cognitive research that inspired the game's design was a more accurate description of learning in the game than other baseline models. Cognitively informed variations on this initial model provided further nuance to our understanding of skill in the game. Finally, an exploration of the differences between our model's predictions and player performance highlighted a previously unforeseen strategy that is potentially distracting students from the goals of the game. Each of these findings was made possible through the application of empirical learning curve analysis and knowledge component modeling.

Now that we have a better understanding of the game the next step is to recommend changes so that *Beanstalk* can further develop players' ability with the concepts of the balance beam. A clear recommendation that follows from our findings is to remove the possibility of the rote solution strategy altogether, or better yet, to keep only the levels on which a rote solution can be constructed but fails as a way of demonstrating the strategy's lack of generality. If players have the ability to fall back on simplistic logic, they are less likely to engage with the game's target concepts to reach a deeper understanding of physics. The possibility for a rote strategy is simple enough to detect with a script, but in more complex cases, other researchers have developed techniques for generating level configurations that are guaranteed to avoid such short cut solutions [40].

A second recommendation follows from the finding that players appear to approach mastery of the concept of mirroring but do not progress quite as far in learning concepts related to conflict-balance, with the one exception highlighted by CB-GMP-ASP levels. One way of interpreting this result is that players clearly get sufficient, and perhaps too much, practice with mirroring while conversely not getting enough practice with conflict balance levels. This would suggest that the pacing of skill ramp up should be altered to get away from mirroring levels sooner in favor of more conflict-balance designs. One potential solution could be to move the CB-GSP-AMP levels to be earlier in the game, because they demonstrate the greatest learning among CB levels, and their reframing of the sum of cross products rule could make the conflict-balance concept more approachable. An alternative approach would be to include more scaffolding for sense making, i.e. helping players to see and reason through the physical process behind the game. This kind of alteration could manifest through tutorial dialogs or extra hinting mechanisms. These changes have the potential to make the game more effective, educationally.

In addition to iterating on the game design we could further refine our modeling procedures to incorporate other processes from the educational data mining literature. For example, in our work, we omitted the "worked example" levels because their being impossible to fail introduced systematic bias into our analysis despite these levels dealing with particular KCs. An elaboration of AFM, called the Instructional Factors Model, has been proposed that is designed to deal with cases when learners are exposed to a concept but do not directly practice it themselves [11]. This expanded model could entertain the possibility that players learn from the exposure to a particular level design, such as a worked example or a tutorial level, without having to play through it to get success feedback. This new model is not yet integrated into the DataShop workflow but it is an approach we plan to explore in future work.

While we have discussed empirical learning curve analysis in the context of an educational game, there is nothing preventing the method from being applied to purely entertainment games. Indeed, the formulation of a learning curve should be applicable to any environment where continued exposure to a concept can be expected to result in better performance [33]. Applying our approach to an entertainment game would require a means of logging player performance with tags for when a skill has the potential to be exercised and whether it was exercised correctly, a process that should be easy with common game telemetry approaches [21,38]. While DataShop provides a number of useful visualization and data management tools, the AFM model can be run relatively simply as a logistic regression, available in most statistical packages.

One nuance of the AFM model is that it assumes that learner's performance trends to an error rate of 0. This is because, in an educational context, AFM is most often applied to data that assumes a model of mastery learning [15]. In an entertainment context it is often desirable that even expert players have some appreciable chance to fail in a game because the risk sustains engagement [16,28]. A further elaboration of the AFM model has recently been proposed, called the Additive Factors Model + Slip, which adds an additional parameter to each KC that allows the performance estimates on KCs to converge to a non-zero value [30]. This elaboration of the model would allow for estimates of initial difficulty for a level as well as a measure of mastered difficulty, or hard the level is even when one has had experience with a concept.

It is worth noting that this type of learning analysis is most appropriate when an objective understanding of skill is required, i.e. when player ability measured is with respect to the task alone and not contingent on other players' ability. The orientation toward objective understanding of skill is driven by the method's original educational purpose. From an educational perspective it is desirable to measure skill within the context of a general task that has the potential to transfer outside of the game, however such transfer must still be validated with measures external to the game. Other models exist for player skill estimation within the subjective context of a player base [19], however these approaches tend not to explicitly account for player learning and they do not provide an understanding of how players are learning the game itself in the absence of other players.

One potential downside of this approach is its requirement of a large base of player data. AFM was originally developed for step-level intelligent tutoring data, which is commonly captured in classroom studies. Since AFM's statistical formulation has one parameter for each student and two for every KC (intercept and slope) it can easily reach a high dimensionality that would be unsuitable to fit with small datasets. This restriction could prevent it from being applied in earlier design stages and initial playtests where the required amount of data is not yet available. In such cases it might be better to apply a more rational design approach such as the one presented by Linehan and colleagues who examined the progression of puzzle introduction in commercial games as learning curves [26]. This approach requires no true player data, and so differs from our empirical learning curve analysis, but can provide insight into difficulty progression in a game and potentially scaffold an eventual empirical analysis in the future.

CONCLUSION

In this paper we demonstrated an application of empirical learning curve analysis to the educational game *Beanstalk*. Through this approach we were able to develop an appropriate model of the skills exercised by the balance beam tasks in the game. Our skill model development was informed by foundational research, rational analysis, and empirical investigation. This new model provided insights into how the game can be redesigned to better accomplish its educational goals and also highlighted a previously unforeseen shortcut strategy to some game levels. This work represents a first demonstration of how these kinds of educational data mining techniques can inform the understanding of skill in games while yielding actionable design recommendations. We hope that other researchers and game design practitioners can benefit from applying similar approaches to their own game, educational or otherwise, and look forward to future work exploring the dynamics of skill acquisition in games.

ACKNOWLEDGMENTS

We would like to thank the designers of *Beanstalk*, our colleagues who performed the study that provided our data, as well as Christopher MacLellan and Kelly Rivers for comments on earlier drafts of this work. This work was supported in part by the DARPA ENGAGE research program under ONR Contract Number N00014-12-C-0284 and by a Graduate Training Grant awarded to Carnegie

Mellon University by the Department of Education #R305B090023. All opinions express in this article are those of the authors and do not necessarily reflect the position of the sponsoring agency.

REFERENCES

1. Akaike, H.A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control 19*, 6 (1974), 716–723.
2. Aleven, V. a W.M.M. and Koedinger, K.R.An effective metacognitive strategy: Learning by doing. *Cognitive Science 26*, 2 (2002), 147–179.
3. Aleven, V., Dow, S., Christel, M., et al.Supporting Social-Emotional Development in Collaborative Inquiry Games for K-3 Science Learning. *Proc. GLS 9.0*, ETC Press (2013), 53–60.
4. Aleven, V. and Koedinger, K.R.Knowledge Component Approaches to Learner Modeling. In R.A. Sottilare, A. Graesser, X. Hu and H. Holden, eds., *Design Recommendations for Intelligent Tutoring Systems: Volume 1 - Learner Modeling*. U.S. Army Resarch Laboratory, 2013, 165–182.
5. Aleven, V., Myers, E., Easterday, M., and Ogan, A.Toward a Framework for the Analysis and Design of Educational Games. *Proc. DiGITEL 2010*, IEEE (2010), 69–76.
6. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., and Qin, Y.An integrated theory of the mind. *Psychological review 111*, 4 (2004), 1036–60.
7. Anderson, J.R. and Lebiere, C.*The atomic components of thought.* Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1998.
8. Baker, E.L., Chung, G.K.W.K., and Delacruz, G.C.The Best and Future Uses of Assessment in Games. In *Technology-Based Assessment for 21st Century Skills*. 2012, 229–248.
9. Baker, R.S.J. d, Habgood, J.M.P., Ainsworth, S.E., and Corbett, A.T.Modeling the Acquisition of Fluent Skill in Educational Action Games. In *User Modeling 2007*. 2007, 17–26.
10. Cen, H., Koedinger, K., and Junker, B.Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. *Proc. ITS 2006*, Springer (2006), 164–175.
11. Chi, M., Koedinger, K.R., Gordon, G., Jordan, P.W., and VanLehn, K.Instructional Factors Analysis: A Cognitive Model For Multiple Instructional Interventions. *Proc. EDM 2011*, (2011), 61–70.
12. Christel, M.G., Stevens, S., Champer, M., et al.Beanstalk : A Unity Game Addressing Balance Principles, Socio-Emotional Learning and Scientific Inquiry. *Proc. Int'l Games Innovation Conf.*, IEEE (2013), 36–39.
13. Clark, D.B., Tanner-Smith, E.E., Killingsworth, S., and Bellamy, S.*Digital Games for Learning: A Systematic Review and Meta-Analysis (Executive Summary).* Menlo Park, CA, 2013.
14. Cleveland, W.S., Grosse, E., and Shyu, W.M.Local Regression Models. In J.M. Chambers and T.J. Hastie, eds., *Statistical Models in S.* Wadsworth & Brooks/Cole, 1992.
15. Corbett, A., Mclaughlin, M., and Scarpinatto, K.C.Modeling student knowledge: Cognitive tutors in high school and college. *User Modelling and User-Adapted Interaction 10*, 2-3 (2000), 81–108.
16. Csikszentmihalyi, M.*Flow the Psychology of Optimal Experience.* Harper Collins, New York, NY, 1990.
17. Gee, J.P.*What Video Games Have to Teach Us About Learning and Literacy.* Palgrave Macmillan, New York, New York, USA, 2007.
18. Heathcote, a, Brown, S., and Mewhort, D.J.The power law repealed: the case for an exponential law of practice. *Psychonomic bulletin & review 7*, 2 (2000), 185–207.
19. Herbrich, R., Minka, T., and Graepel, T.TrueSkill^{TM}: A Bayesian Skill Rating System. *Nips*, (2007), 569–576.
20. Iacovides, I., Cox, A.L., Avakian, A., and Knoll, T.Player Strategies : Achieving Breakthroughs and Progressing in Single-player and Cooperative Games. *Proc. CHIPLAY 2014*, ACM Press (2014), 131–140.
21. Kim, J.H., Gunn, D. V, Schuh, E., Phillips, B.C., Pagulayan, R.J., and Wixon, D.Tracking Real-Time User Experience (TRUE): A comprehensive instrumentation solution for complex systems. *Proc. CHI 2008*, ACM Press (2008), 443–451.
22. Koedinger, K.R., Baker, R.S.J. d, Cunningham, K., Skogsholm, A., Leber, B., and Stamper, J.A Data Repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy and R.S.J. d. Baker, eds., *Handbook of Educational Data Mining*. 2010, 43–55.
23. Koedinger, K.R., Corbett, A.T., and Perfetti, C.The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science 36*, 5 (2012), 757–98.
24. Koster, R.*A Theory of Fun.* Paraglyph Press, Inc., Scottsdale, AZ, 2005.
25. Landau, L.Pyschometrica Considerations in Game-Based Assessments. *White Paper Released by Glasslab*, (2014), 160.

26. Linehan, C., Bellord, G., Kirman, B., Morford, Z.H., Roche, B., and Kildare, C.Learning Curves : Analysing Pace and Challenge in Four Successful Puzzle Games. *Proc. CHIPLAY 2014*, ACM Press (2014), 181–190.

27. Linehan, C., Kirman, B., Lawson, S., and Chan, G.G.Practical , Appropriate , Empirically-Validated Guidelines for Designing Educational Games. *Proc. CHI 2011*, ACM Press (2011), 1979–1988.

28. Lomas, D., Patel, K., Forlizzi, J.L., and Koedinger, K.R.Optimizing challenge in an educational game using large-scale design experiments. *Proc. CHI 2013*, ACM Press (2013), 89–98.

29. Long, Y. and Aleven, V.Gamification of Joint Student/System Control Over Problem Selection in a Linear Equation Tutor. *Proc. ITS 2014*, Springer (2014), 378–387.

30. Maclellan, C.J., Liu, R., and Koedinger, K.R.Accounting for Slipping and Other False Negatives in Logistic Models of Student Learning. *Proc. EDM 2015*, (2015), 53–60.

31. Malone, T. w.Toward a Theory of Intrinsically Motivating Instruction. *Cognitive Science 4*, (1981), 333–369.

32. Manske, M. and Conati, C.Modelling Learning in an Educational Game. In C.-K. Looi, M. Gord, B. Bredeweg and J. Breuker, eds., *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*. IOS Press, Amsterdam, Netherlands, 2005, 411–418.

33. Martin, B., Mitrovic, A., Koedinger, K.R., and Mathan, S.Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction 21*, 2011, 249–283.

34. Newell, A. and Rosenbloom, P.S.Mechanisms of skill acquisition and the law of practice. In J.R. Anderson, ed., *Cognitive skills and their acquisition*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1981, 1–56.

35. Rasch, G.*Probabilisitic Models for Some Intelligence and Attainment Tests.* MESA Press, Chicago, IL, USA, 1993.

36. Schell, J.*The Art of Game Design: A Book of Lenses.* Morgan Kaufmann, Burlington, MA, 2008.

37. Schwarz, G.Estimating the Dimension of a Model. *The Annals of Statistics 6*, 2 (1978), 461–464.

38. Seif El-Nasr, M., Drachen, A., and Canossa, A., eds.*Game Analytics.* Springer London, London, 2013.

39. Siegler, R.S.Three Aspects of Cognitive Development. *Cognitive Psychology 8*, (1976), 481–520.

40. Smith, A.M., Butler, E., and Popovi, Z.Quantifying over Play: Constraining Undesirable Solutions in Puzzle Design. *Proc. FDG 2013*, (2013), 221–228.

41. Stamper, J.C. and Koedinger, K.R.Human-machine student model discovery and improvement using Data. *Proc. AIED 2011*, Springer (2011), 353–360.