A Human-Centered Approach to Data Driven Iterative Course Improvement

Steven Moore[1(✉)], John Stamper[1], Norman Bier[1], and Mary Jean Blink[2]

[1] HCII, Carnegie Mellon University, Pittsburgh, USA
stevenjamesmoore@gmail.com, jstamper@cs.cmu.edu, nbier@cmu.edu
[2] TutorGen, Inc., Wexford, USA
mjblink@tutorgen.com

Abstract. In this paper we show how we can utilize human-guided machine learning techniques coupled with a learning science practitioner interface (DataShop) to identify potential improvements to existing educational technology. Specifically, we provide an interface for the classification of underlying Knowledge Components (KCs) to better model student learning. The configurable interface allows users to quickly and accurately identify areas of improvement based on the analysis of learning curves. We present two cases where the interface and accompanying methods have been applied in the domains of geometry and psychology to improve upon existing student models. Both cases present outcomes of better models that more closely model student learning. We reflect on how to iterate upon the educational technology used for the respective courses based on these better models and further opportunities for utilizing the system to other domains, such as computing principles.

Keywords: Learning analytics · Student model · Learning curve · Data visualization · Data-driven improvement · Educational technology

1 Introduction

The proliferation of data on students interacting with online learning environments has opened up enormous possibilities for understanding student behavior for decades [1]. It enables the construction of models on how students progress through the learning process and identify the gaps in their knowledge. Building on these student models for the purpose of tracking student learning over time has been a key area of focus in the educational technology community [2]. Cognitive tutors, such as those from Carnegie Learning, utilize student models and are adaptive to student knowledge by tracking the mastery of skills or knowledge components (KCs) [3]. The models that map KCs are generally created with the help of subject matter experts and cognitive scientists. Unfortunately, these knowledge component models (KCMs) do not always correctly model skills, which can impede student learning. When a KCM for a cognitive tutor is incorrectly modeled, it can cause incorrect problem selection and waste valuable student time on skills they have

M. E. Auer and D. May (Eds.): REV 2020, AISC 1231, pp. 742–761, 2021.
https://doi.org/10.1007/978-3-030-52575-0_61

already mastered. While it is challenging to get the models perfect, continuously iterating on the model as more data is collected can help to improve it.

Learning analytics can address this problem and presents an opportunity for continuous improvement of the models using data driven techniques [4]. In this paper, we show how we can use new user interface affordances in DataShop [5], that utilizes a novel framework for curve categorization, to assist in identifying areas of improvements in the student models of the educational technology. Using the curve categorizations as a starting points, novice users are able to make model improvements using the affordances of DataShop. We present two case studies in different educational technology systems across unique educational domains where the new DataShop features were used to improve the underlying KCM.

2 Related Work

2.1 Knowledge Components

When a student is solving a problem, there are a series of hypothesized competencies, or knowledge, that are needed to perform each step of the problem. These competencies are known as knowledge components (KCs). The KCs are fine grained representations of knowledge that includes constraints, schemas, and production rules [4]. In an educational technology system, such as a cognitive tutor, each of these problem steps has an associated action the student needs to take to solve the problem [6]. These actions can be labeled with one or more KCs to represent the required competency a student needs to successfully solve that step. Each of these problem steps also corresponds to an opportunity at which a student must demonstrate their mastery with the mapped KCs.

While it may have once been question *if* student learning could accurately be modeled through their progression of KCs, evidence has shown this to be the case for a variety of domains using knowledge tracing [7, 8]. Knowledge tracing is the practice of estimating a student's current knowledge state at a given time while they interact with an educational technology system [9]. It maintains a record of the probability that a student knows a skill or concept, based on their performance on problem steps that are mapped to KCs. This is used to inform different functions of the ed tech system, such as problem selection or advancing the student to a new content area, based on their current mastery level with the KCs. It is important to have an accurate KCM for this reason, as a poorly fit one can lead to inaccurate knowledge tracing, where over-practice or under-practice may occur [10].

2.2 Learning Curves

A learning curve is a graphical representation of the change in student performance over time. Learning curves show where students begin with their knowledge, the rate at which they learn the given KCs, and the flexibility of how the acquired skills can be used [11]. In DataShop, learning curves have the opportunities a student has with a particular KC on the x-axis and the error rate, as a percent, on the y-axis. By default,

each learning curve corresponds to a single KC in DataShop and ideally, the number of students with an error, represented at each opportunity point, decreases. This decrease in students at each point occurs for a variety of reasons, such as students not attempting a problem again or skipping activities that are mapped with that KC. We know students learn more by doing, such solving problems in a course rather than just reading the text [12]. Therefore, it is expected to see a downward sloping learning curve with each additional opportunity to practice. As shown in Fig. 1, a desired learning curve is one that shows student improvement, the error rate decreasing, as the opportunity count with the specific KC increases. Additionally, these learning curves are reflective of the effectiveness of a learning system and its content, as it shows learning for the group of students using the system.

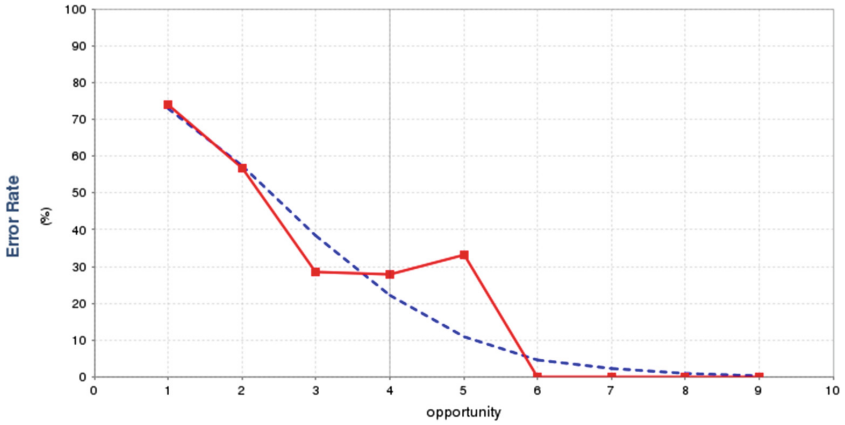

Fig. 1. A learning curve categorized as *Good*, showing a decreasing error rate per opportunity

The analysis of learning curves to provide insights into student models has been around for many years [13]. Methods involving manual human inspects of the curves to more semi-automated ones have been used to improve upon cognitive models used in intelligent tutoring systems [14]. Further analysis of learning curves and their categorizations to provide insights into courseware via crowdworkers has also been suggested [15]. Learning curves in DataShop use a specialized form of logistic regression performed on the error rate of the curve, known as the Additive Factors Model (AFM) [10]. AFM is a statistical model that makes predictions regarding student performance, in combination with item response theory that includes a growth term [16]. This model uses information about a student's prior practice opportunities on the assigned KC to predict the probability that the student will perform correctly on a given opportunity, which corresponds to a question step.

2.3 Prior DataShop Work

Previous research around DataShop has focused on improving the underlying KCMs in order to gain insights into the student learning process and provide suggestions for

educational technology improvement [4, 17, 18]. Work by Stamper and Koedinger [4] showed how human-centered aspects could be combined with DataShop tools using a human-machine discovery approach to improve student models. This human aspect comes into play to make distinctions involving learner populations, sequencing, and mis-tagging of KCS that the machine learning side might miss or not take into account. For instance, teaching the same algebra course to middle vs. high school students would result in the same dataset, but it should be split amount these student populations, something the human would have knowledge of and do. The improved student model from this method was applied to two other datasets in the same domain, but collected from different sets of students, and still demonstrated improvements. Building upon that, one study utilized AI and statistical methods to discover improved models in a variety of domains using data collected from different educational technologies [18]. These resulting improved models isolated flaws in the original one, for which they demonstrated how an investigation on these flawed parts led suggested improvements for tutor design. Using these methods, a recent study was able to improve the KC model for an educational math game [19]. The authors were able to better identify parts of the game that gave the students trouble and make improvements in the form of question content and ordering.

These studies led to later work focused on an initial close-the-loop experiment, where the improved KCMs were tested inside the classroom. The results showed gains of 25% less time to master the same material and improved performance on a subset of problems using a particular skill in the course, by using the improved models in the tutor [17]. A recent study analyzed KCMs in DataShop by comparing and contrasting their metrics to measure how accurately the predictive fit of the models is to the data. They found that of the metrics, Akaike information criterion (AIC) was the best predictor of cross validation results, which is the gold-standard for model selection [20]. Another metric relevant to student models is Bayesian information criterion (BIC), while similar to AIC, denotes how well the AFM statistical model fits the data with the given KCM. In addition to these, root mean square error (RMSE) is often used, which shows how well the KCM might generalize to an independent dataset from the same educational technology, such as a cognitive tutor. Much research has focused on improving KCMs, leading us to build upon that work to find a solution for easy categorization of learning curves. These categorizations can contribute to advancing learning following a human-centered approach using data-driven iterations.

3 DataShop

3.1 Functionality

The Pittsburgh Science of Learning Center DataShop (pslcdatashop.org) is the world's largest repository of learning interaction data for research [21]. It provides a suite of tools for researchers, instructional designers, and data scientists to analyze, create, and modify educational data. Student log data from many educational technology systems are fed into DataShop, where the student interactions with the questions in the given platform can then be analyzed. The analysis features include viewing the KCs

associated with the questions, viewing their problem statements, associated learning curves, and more detailed statistics, such as student accuracy and how many times the question was attempted. This collected data is fine-grained, with an average student action logged every twenty seconds. Additionally, this data is often longitudinal, spanning courses that are half a semester, a full semester, or even a yearlong.

DataShop offers statistics and categorizations on learning curves and knowledge component models. The built-in suite of tools provides a way to view the details on learning curves, including categorizations of them, to help visualize how student learning changes over time for a particular KC [22]. We focus our case studies on the use of these categorizations, showing they can be used by non-experts to make judgements about KCs to further analyze. These categories for the learning curves are driven by a set of configurable parameters found on DataShop's interface for viewing the curves. The parameters, shown in Table 1, may vary based on the data set that is being viewed. A user might modify the threshold for students to a value below ten if the class from which the data were gathered is small. Similarly, if each knowledge component is only assessed a few times, then the opportunity threshold might also be reduced. These are two parameters where knowledge of the student population from which the data was generated from would inform how a user sets these.

Both the low and high error thresholds might be tweaked depending on the domain or prior knowledge of the students. Finally, the AFM slope threshold might change based on the rate of learning students are expected to improve upon. For instance, if they are expected to reach mastery after several problems, this threshold might be lowered. If the curve does not violate any of these parameter thresholds, then it is categorized as *Good*. Ultimately the desired value of the parameters will vary based on the system that collected the data. For instance, if the tutor only contains a few questions per KC, intended to be used as review, the opportunity and error thresholds would need to be tweaked to get a better categorization of the data.

Table 1. Configurable parameters for learning curve categorization

Parameter threshold	Description	Default value
Student	The minimum number of students that have attempted the KC at each point in the curve	10.0
Opportunity	The minimum number of student attempts at a KC that must be present for a curve, if the attempts are below this value the curve is labeled too little data	3
Low error	If a point on the curve falls below this error rate value, the curve is low and flat	20.00
High error	If the last point on the curve goes above this error rate value, the curve is still high	40.00
AFM slope	If the calculated AFM slope of the curve falls below this value, the curve is labeled as no learning	0.001

3.2 Data Sources

A primary educational technology platform that feeds data into DataShop is the Open Learning Initiative (OLI). OLI is an open educational resources project, part of the Simon Initiative at Carnegie Mellon University, that allows instructors to develop and deliver online courses consisting of interactive activities. Detailed student interactions with the course materials, such as watching videos, answering a variety of traditional question types such as drag-and-drop, multiple choice, and responding to free-form question prompts are logged into DataShop. Each question in OLI is broken down into one or more problem steps, where each step corresponds to an opportunity, the x-axis for a learning curve in DataShop. For instance, if a question asks a student to set the value of three dropdown boxes, then that question has three steps. In addition to the traditional timestamps and UI element with which the student interacted, each step is assigned a set of one or more KCs required by the student to answer the question. This KC tagging of the questions in conjunction with student accuracy on the problem, time of task, and number of attempts, provides detailed insights into which concepts with which students are struggling most.

Another platform that feeds a large portion of data into DataShop is Carnegie Learning's MATHia, a cognitive tutor for algebra (carnegielearning.com). These tutors cover middle and high school math curriculum and are adaptive. Students work math problems that feature rich interactions and these interactions are logged at every step. All steps are tagged with knowledge components and associated KCMs are also exported to DataShop. With a detailed log file imported into DataShop, we can use this data to track learning over time and perform learning curves analysis. Carnegie Learning data was also featured in the 2010 KDD Cup data mining competition [5] hosted by DataShop (pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp).

3.3 Curve Categorization

The categorization of learning curves feature was added to DataShop following insights from prior research on improving student models [4]. Learning curves are now automatically grouped into one of five categories, based on the parameters. This categorization is a hierarchical approach, based on the configuration variables noted for the given dataset. They provide a way for users to gain a better grouping and view of the learning curves for their data, based on the parameters the user can alter to better fit the context of the educational system from which the data was collected. The categorization of the curves first accounts for the opportunity threshold, thus categorizing any curve with fewer opportunities than that parameter as *Too Little Data*. Following this, the second parameter considered is the student threshold, which is the minimum number of students that have attempted a KC at each point in the given curve. The category formulas only use points on the curve that are at or above the student threshold, which defaults to ten. This means that a curve's final points may not be utilized in the categorization due to having too few student observations.

The first category is *Low and Flat*, which indicates that students have already mastered the target KC and do not need the additional practice. Curves in this category begin with a very low error rate and remain low as the opportunity count progresses, as

shown in Fig. 2. For KCs that are grouped into this category, it is suggested that the number of assessments targeting it be reduced to avoid over-practicing [10]. A student's time is better allocated toward a set of different KCs for which they have yet to achieve mastery. When an intelligent tutoring system is the educational technology used for a dataset with curves in this category, it may also be the case that the knowledge-tracing parameters are misaligned, and the system is suggesting further practice that is redundant. It may also be the case that a different system has too many practice problems for a particular KC and that they should remove some of them in favor of other material.

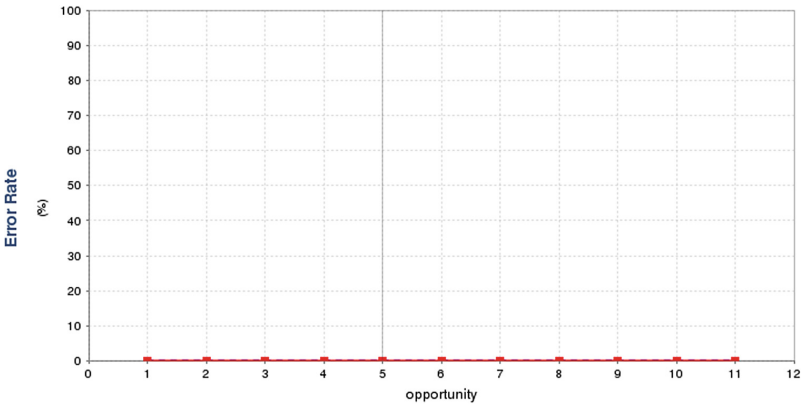

Fig. 2. *Low and Flat* learning curve showing students starting at 0% error rate and receiving up to 11 assessments mapped to this KC

Similarly, the second category, *No Learning*, is a curve representative of student learning where they do not demonstrate learning gains at a significant level. Curves in this category often begin at a moderate error rate and end around the same rate for the fitted curve, represented by the dashed line, shown in Fig. 3. These occur when the predicted learning curve's slope does not show apparent learning for the given KC. Even when the curve's final point is above the high error threshold, this category takes priority over the subsequent ones based on its use of the AFM slope in categorization. It is also important to remember these types of curves are potential cases to explore breaking down the KC into multiple KCs or disaggregated based on student subpopulations, as previous work has shown prior to the implementation of categorizations [4, 17, 22, 23].

The *Still High* category is another type of curve that is an easy area where potential improvements could be made to the KCM. Learning curves in this category have their final point, that is at or above the student threshold, above the high error threshold parameter value. This indicates that students continued to struggle with any KCs in this category, despite having sufficient opportunities. It is recommended that these curves be analyzed for another potential case of breaking down a single KC into multiple ones or providing students with additional practice opportunities. For instance, it may suggest that a better intermixing of practice be done, such as reviewing worked examples and then solving problems [24]. Figure 4 demonstrates how even when

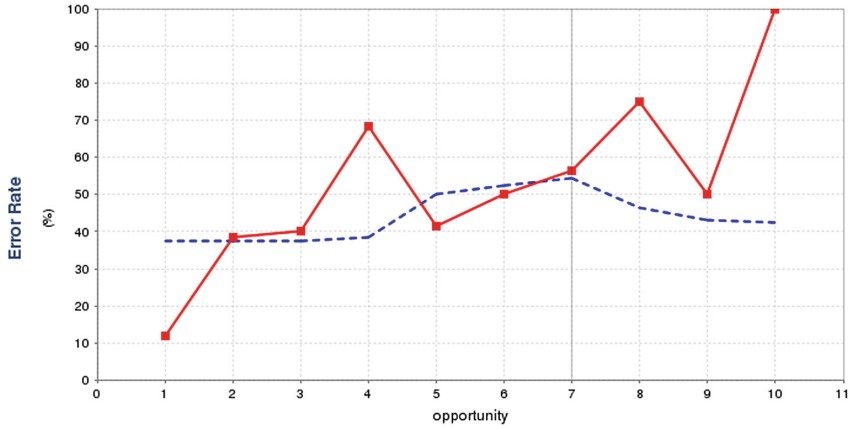

Fig. 3. *No Learning* curve where the predicted learning is below the AFM slope

students demonstrated learning, decreasing their error rate as they have opportunities, the curves end points may still fall above the set high error threshold, which was at the default value of forty.

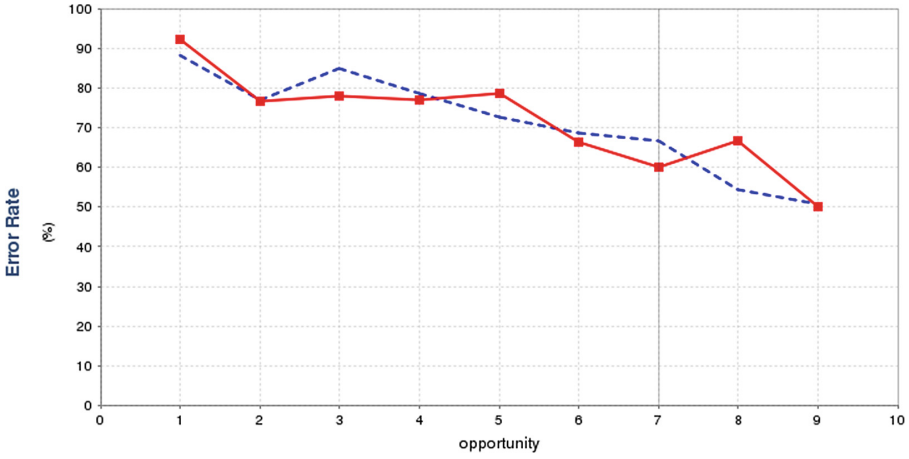

Fig. 4. *Still High* learning curve with the final point above the high error threshold (40)

When students did not have enough practice opportunities with the KCs, they are categorized as *Too Little Data,* since there are not enough opportunities for the data to be meaningful based on the configured parameters. These curves are based on the opportunity threshold and curves below the configured value are categorized as such. Even when a curve may show points above this opportunity value, the formula for generating the curves, using AFM, only includes points that meet or exceed the student threshold. Thus, by default curves with three or less opportunities of ten or more

students are grouped into this category (see Fig. 5). It is recommended that more practice opportunities be added for these KCs, so they can be assessed with enough points to determine student learning progress.

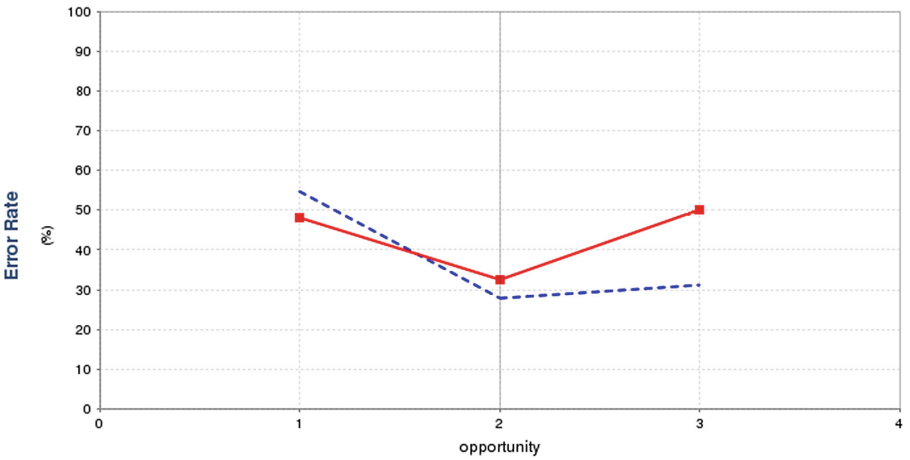

Fig. 5. Learning curve with only three points, categorized as *Too Little Data*

Finally, learning curves that did not get categorized into the aforementioned "at risk" ones are labeled as *Good*. The previous curves are "at risk" ones due to their being an opportunity for improvement in them. However, curves in the *Good* category indicate that student learning is occurring as they progress through corresponding assessments. While curves of this nature may still have room for improvement, these have an optimal balance of student improvement as opportunity count increases and are

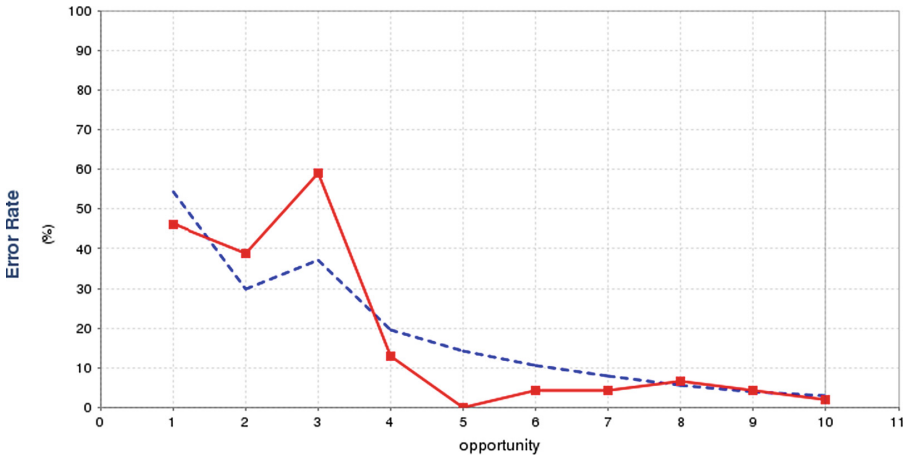

Fig. 6. Learning curve categorized as *Good* using the default parameter values

less likely to have easily identifiable areas of breakdown. A curve can still be Good even if all the points do not decrease in error rate, as demonstrated in Fig. 6.

4 Evaluation

4.1 Case Study One - Psychology

When interpreting learning curves in DataShop, one of the underlying assumptions is that the existing KC model is accurate. For this case, which models the way a student learns, we define accurate as representing valid segmentation and progression of the content. After all, it is often the case that the model in question was constructed by an expert of that particular domain who has an intuitive understanding of the content. What is important to consider, however, is that decisions regarding the KC model can be primarily driven by discipline specific standards, rather than by an iterative, data-informed process that is more broadly supported by learning sciences. When we examine the learning curves, these considerations allow us to identify candidates for human intervention to make appropriate changes so that we have a better fitting model. Additionally, by leveraging other users in the analysis and refinement process, we can hope to avoid some expert blind spot that may have been present in the generation of the original model by the domain expert.

To begin, we started our analysis of a Psychology dataset in DataShop recorded from an OLI course. The user for this case study had never used DataShop previously, but wanted to investigate a Psychology course dataset as part of their learning and participation in an educational data mining workshop. This dataset was made available to them, as it contains recent data from being used the past few years by students as part of their university class in Psychology. This data consisted of logs from 180 students who took the course during one of two semesters. When we first began analyzing the data, there were 272 total KCs. Due to the volume of this dataset, the student threshold parameter was set to 20 and the opportunity threshold was set to 5, so the data displayed in the curves would have more student attempts and a great number of assessments. Modifying these parameters changed how the learning curves for this dataset were grouped.

We were most interested in examining learning curves that have been categorized as *No Learning* or *Still High*. These curves showed plenty of room for improvement and it is common for curves with alignment issues to end up in these categories. Curves with alignment issues are ones that have problems mapped to KCs that are poor fitting and not representative of what is required to solve the problem. We first noticed a learning curve for a particular KC categorized as *No Learning*, "describe_psyhcoactive_drugs", that begins to descend normally and then spikes suddenly, shown in Fig. 7. We interpret the curve as telling us that at first, the students are learning predictably, making fewer and fewer errors. When the curve spikes, it is an indication that students have suddenly begun making errors at a much higher rate, which is confounding given the initial learning progress indicated by the beginning of the curve. To determine the cause of the increased error rate, we examined the individual corresponding items in the lesson that were mapped with this KC.

Fig. 7. Learning curve, categorized as *No Learning*, with a suspicious spike at opportunity 19

Using domain expertise, judgements were made regarding whether those opportunity items were constructed appropriately or categorized correctly. It appeared that the problems at and after the spike assessed different knowledge than prior items on the curve. The first half of the learning curve was from problems describing a particular type of psychoactive drug, while the latter part had questions about a different type of drug. As a result, we split this KC into two KCs since there were enough opportunities to provide sufficient results for each. The resulting two KCs still had enough opportunities for the given threshold parameter to be analyzed. The resulting curves, Fig. 8 and Fig. 9, show much smoother learning curves than the original one for the given KC and are now categorized as *Good*.

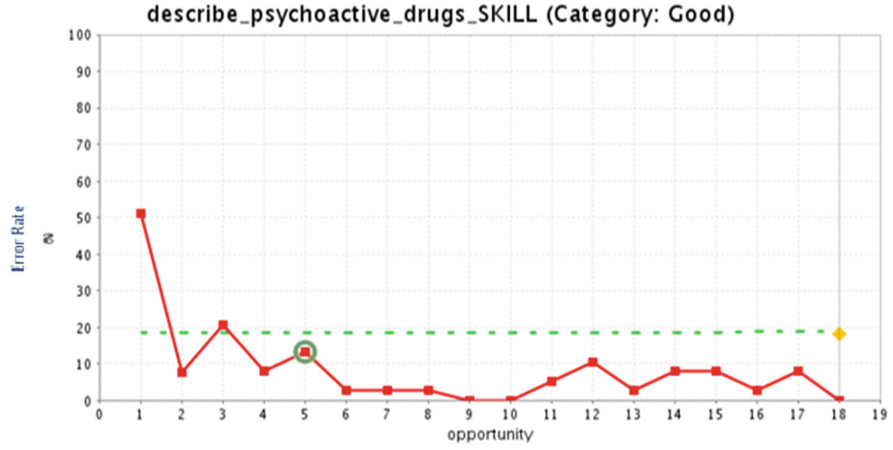

Fig. 8. The first half of the original describe_psychoactive_drugs learning curve, with the points at opportunities 19 and beyond remapped onto a different KC

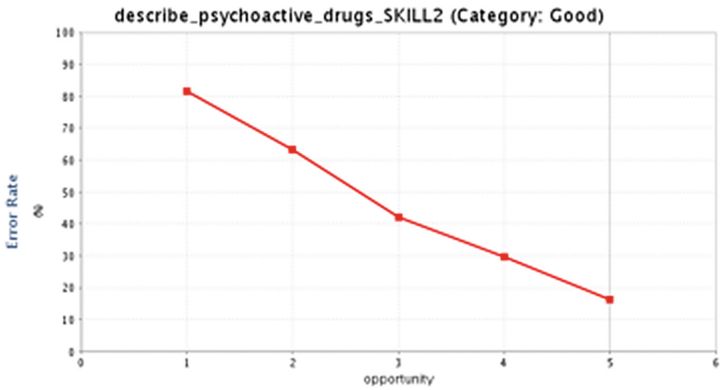

Fig. 9. The learning curve for the added KC, split from the original describe_psychoactive_drug KC

Another similar occurrence of a suspicious spike was seen when analyzing the curves categorized as *Good*. Despite this spike, students were learning this KC as the error rate was decreasing at an appropriate rate. While there were other curves to investigate, we decided to look at this one due to it have a spike like the previous one. A learning curve shown in Fig. 10, "analyze_intelligence_concepts", appeared to have a spike around the tenth and eleventh opportunities. Upon inspecting the questions in the course tagged with this KC, we noticed two of the questions were not assessing analysis of intelligence concepts. The two problems were having students identify and define, which was a better fit for an existing KC titled "describe_intelligence_concepts". Assigning those two opportunities to this different KC yielded a smoother curve and better assessed the appropriate skills. As a result, both the "analyze_intelligence_concepts" and "describe_intelligence_concepts" were improved, showing a smoother curve, as a result of this re-tagging.

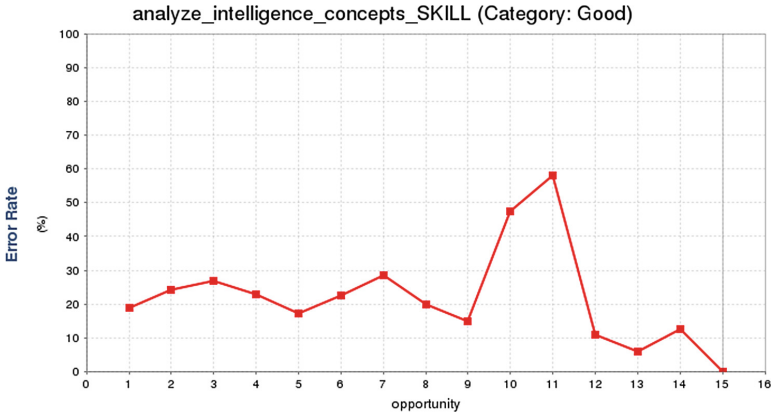

Fig. 10. A *Good* learning curve for analyze_intelligence_concepts that has an area of spiking at opportunities 10 and 11

The final learning curve we looked at was one for the KC "identify_therapies_-modalities_practices" that was categorized under *No Learning*, shown in Fig. 11. This curve had a suspicious spike, like the previously investigated learning curve, at one opportunity we had been looking for, but also had another spike consisting of three opportunities later in the curve. We analyzed problems mapped with this KC that occurred at the spike and, through our case study user's domain knowledge, determined they were indicative of research methods knowledge rather than identifying therapeutic modalities. An example of one such question, better suited for a different KC than what it was originally mapped to, is shown in Fig. 12. Remapping the problems at the spike with their more fitting research methods KC made our original curve smoother and removed the first spike. However, the later region of spiking became more pronounced, due to the fewer number of students who had done a problem at that opportunity count. Students who had not demonstrated mastery were continuing to try the problems, increasing the overall error percentage at that point. While this later spike in the curve might not warrant a remapping, it does suggest that the content might need to be improved so that students can hopefully achieve mastery before reaching this many attempts.

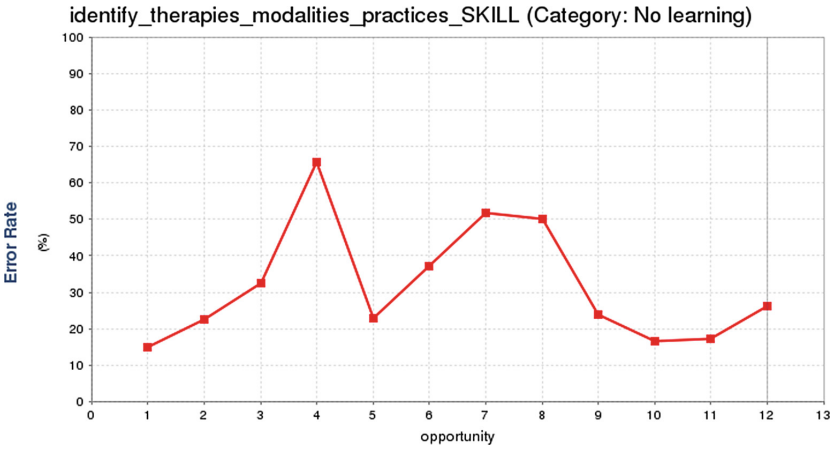

Fig. 11. Learning curve for identify_therapies_modalities_practices that has an initial spike at opportunity 4 and another set of spikes around opportunity 7

1) In an effort to monitor the effectiveness of different therapies and other medical treatments, studies are conducted. The results of such studies are published and are otherwise known as -- ▾ (research outcomes)

Fig. 12. A question used in the course that was originally mapped to the KC for identifying therapies, but is more fitting for the research methods KC

4.2 Case Study One - Model Improvement Results

To validate the hypothesized model improvements for both cases, we performed a parallel analysis on the original student models compared to revised models with added KCs and re-tagged problem steps. The original student model for the first Psychology case, "pyschology_1-6", was created by the course's instructor, an expert in the domain. This original model consisted of 272 KCs and after the remapping and decomposition of the user in case study one, a new "psychology_1-6_model4" was developed consisting of four additional KCs for a total of 276. Utilizing AFM, we found the newer model is a better predictor of student learning when compared to the original mode, summarized in Table 2. The KC adjustment led to reducing AIC (176,705 to 176,441), BIC (183,912 to 183,728) and unstratified root mean square error (RMSE) on test set fit in cross validation (0.435319 to 0.435038). These model values support the addition of the KCs and demonstrate how the model can show improvements from just modifying a few learning curves. Not only can these modifications improve the predictive accuracy the model provides, but the analysis provides key human insights into the content that otherwise might be neglected. For instance, the analysis and improvement of these learning curves also allows users to look at the problem associated with these KCs, which might be indicative of refinement for the content, not just the KC associations.

Table 2. Knowledge Component model values for the Psychology course

Model	AIC	BIC	RMSE
psychology_1-6_model4	176,441	183,728	0.448741
psychology_1-6	176,705	183,912	0.449876

4.3 Case Study Two - Geometry

The data used in our second case study is from a Carnegie Learning cognitive tutor unit on Geometry area. This particular unit occurs later in the curriculum, and by the time students reach this unit, the skills around finding shape areas have been merged to a single skill called "Find Individual Area." Earlier units expressly break this skill into multiple skills; one skill for each shape type. By this unit, however, it is expected that students have successfully mastered these skills and now are just addressing find area as plugging in the inputs to the correct area formula for the shape given. The data largely backs up this merging of the individual area skills, but categorization in DataShop still pointed to a potential improvement of the model on this skill.

For this case study, the primary user was a Master's of Data Science student, who chose to user DataShop, for the first time, as part of a class project. When looking at the "Find Individual Area" skill using the KCM that was provided with the dataset that was used in the cognitive tutor, KTracedSkills, we see that it is categorized as *No Learning* as seen in Fig. 13. Further, visual inspection, clearly shows a spike in the error rate at opportunity 5. This led us to explore what problem steps were attempted at opportunity 5 versus the previous opportunities, which seem to show a declining curve. Sure

enough, the majority of problem steps with errors were all centered on problems containing trapezoids. By retagging the "Find Individual Area" skill in all problems that address trapezoids as "Find Trapezoid Area," we were able to get a better fitting model that listed both skills in the *Good* category. This suggests that at this point in the tutor, for the group of students working, that the representation of "Find Individual Area" as 2 separate skills is a better representation of actual student knowledge. Students in the tutor were not at the same level with their skill for other shapes collectively compared to trapezoids. Making this change in the model should lead to improved learning outcomes if the model would be updated.

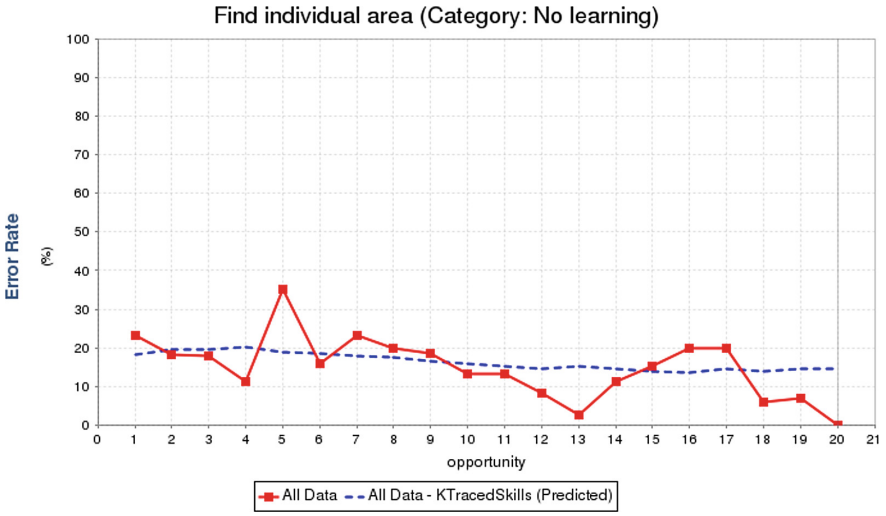

Fig. 13. Learning Curve for a KC called Find individual area that is categorized as *No Learning*. Visual inspection seems to show a potential improvement around opportunity 5, where there is a spike in error rate. Drilling down into the steps in this opportunity show that many of the errors are in problems with trapezoids, suggesting there could be different KC around these problems

4.4 Case Study Two - Model Improvements Results

The original student model for this Geometry case was one used in the cognitive tutor and was provided with the dataset, KTracedSkills. This original model contains 10 KCs and after the addition of the "Find Trapezoid Area" KC, the new model contains a total of 11. Again, using AFM we found the newer model, KTracedSkills-trap, was more predictive of the student data than the original model, summarized in Table 3. The single KC adjustment led to reducing AIC (3,409 to 3,377), BIC (4,215 to 4,196), and unstratified RMSE on test set fit in cross validation (0.304451 to 0.303349). While the improvements weren't as big as the Psychology example, there was still a noticeable difference by modifying a single KC.

Table 3. Knowledge component model values for the Geometry course

Model	AIC	BIC	RMSE
KTracedSkills-trap	3,377.25	4,196.49	0.315229
KTracedSkills	3,409.26	4,215.53	0.317503

5 Discussion

The categorization of learning curves in DataShop provides an initial grouping that allows users, regardless of expertise level, to focus their analysis of the KCs that might benefit the most from refinements. DataShop provides users the ability to filter through hundreds of KCs for their datasets, and analyze which ones are effective or which ones indicate students are not learning at an expected rate. The configurable parameters allow users to filter out learning curves that might not be as relevant for analysis, such as ones with little data, and focus on the more pertinent ones. These parameters also influence how the categorization of curves are formed, by modifying the thresholds, and allows the user a greater level of control. This presents an easy way to drill down into the learning curve to view the student error rate at each opportunity, and find anomalies, such as the high error rate spikes. We showed in our two case studies that this method of categorization to classify learning curves can accurately identify curves deserving further inspection. The analysis performed was able to determine potential issues in the model and make appropriate refinements to it.

This categorization provides a high-level view of all the curves in a manner that suggests which ones should likely be addressed first. The main analyst in the first case study regarding Psychology was a user, with a PhD in Social Psychology, who was new to DataShop. For the second case study, the main analyst was a Master's of Data Science student, also using DataShop for the first time. While these users might be familiar with general data science and statistical practices, the use of DataShop and involvement with learning curves was new to them. They were able to effectively utilize the learning curve categorizations to guide their selections into digging deeper into their analysis. These users utilized other features in DataShop as part of their investigation, but the initial grouping and visual display of all the curves served as the starting point for their analysis.

By improving, splitting, and modifying just a few learning curves, we were able to create a better fitting student model for both cases. The AIC and BIC decreased in each instance for the newer models, meaning that the AFM statistical model fits the data closer, providing a more accurate measure of student learning and progression [8, 16]. Additionally, the RMSE also decreased, suggesting that the new models will generalize better to datasets of that domain from the same tutor. This translates to having increased accuracy from a knowledge tracing perspective, which is important for intelligent tutoring systems. The improved accuracy will help provide the students the correct amount of problems needed to achieve mastery for a given skill, particularly when this process and improvement is applied to multiple KCs. Having this closer fitting model is key in order to avoid over or under-practice [10]. With a more accurate problem

selection, these tutoring systems can help students learn more efficiently and make better use of their limited time. It also models the student learning process better, allowing for the suggestions of next problems that contain only the KCs a student still needs to master and letting them advance through the tutor at the correct pace.

While the Geometry case did not show as much improvement as the Psychology, only a single KC was broken down in that instance. The Geometry dataset came from an ITS used in production, one developed by professionals at a company rather than by a single professor, so the quality of the original model may have been stronger, needing less refinement. This means there may not be as many improvements possible to the model. It is not to say that breaking KCs into multiple ones always leads to an improved model, sometimes the remapping of a KC is required rather than creating a new one, as was the case for the second learning curve in the Psychology example. However, in both studies, only working with a few learning curves led to changes that created new improved models. While we did not get to feed these models back into the host educational technology, OLI and an ITS, several studies support that the improved AIC, BIC, and RMSE scores will result in improved student learning [4, 8, 17]. This is particularly useful for the Geometry case, where the model is used for a tutoring system with knowledge tracing.

Aside from improving the student model, iterative improvements to the educational technology systems are other potential outcomes of such analysis. In addition to adjusting the KCMs, there may be a need to adjust the systems and content to support these model refinements, in order to fully realize the improvements. The data from the Psychology example utilizing OLI had its activities mapped with the original model's KCs by a domain expert, the course instructor. Analysis of the "describe_psychoactive_drugs" learning curve that led to it being broken down two component ones might suggest that new assessments be added for the added component skill. In that case, the original KC still had 18 opportunities, but the added one only had 5. Providing more opportunities for the later could provide a more accurate measure of student learning for that KC and ample opportunities to develop mastery. Similarly, the analysis of the "analyze_intelligence_concepts" learning curve suggested it was assessing a different KC and needed to be remapped. It may be the case that other assessments in the course are actually targeting "research methods" like this KC, but are also mislabeled.

6 Conclusion and Future Work

We presented two cases where novice users of DataShop were able to utilize its features and the categorization of learning curves to assist in identifying potential problem areas within a course. The results of utilizing the learning curve categorization, drilling into the learning curves, and breaking them down into multiple KCs or remapping them, led to improved student models for both cases. Data that created the curves came from two different educational technology systems, yet both benefited from similar methods that utilized the affordances of DataShop. Not only were the technology systems different, but they also represented two completely different domains. However, we were able to apply similar techniques to both in order to improve their corresponding student models. Our study is another step toward showing

how novice users can analyze the large amount of data their educational technology systems collect in a way that feeds into the iterative improvement of courses. It supports that by using the learning curve categorizations as a starting point, users can make informed judgements when it comes analyzing KCs. The improved student models these KCs feed into not only better model learning, but can be used to accurately inform course instructors of their students' learning and areas they might target for course improvement.

Continued work should look at applying similar techniques and utilizing the categorizations to find areas of course improvement in even more diverse domains. One such domain we are moving towards is computing principles, which currently has several years worth of data available in DataShop. Such courses often have a mix of questions types, from programming activities to free response. We believe that analysis of the curves for that datasets will reveal the need for similar interventions as the two presented case studies.

Additionally, this process supported by DataShop offers the potential to create an improved model from a semester's worth of data and see how it translates to many other datasets from different semesters of the same course. Finding how generalizable an improved model is suggested by the RMSE, is important in creating a solution that is effective across all student populations. Additionally, future work should look to feed the improved student models back into the educational technology and measure the learning gains students have from the better fitting model. This is key for intelligent tutoring systems or other educational technology systems that utilize knowledge tracing, as the student model is core to the system. Building upon this, there may be a benefit for looking at ways to help users identify when to remap problems steps to KCs, or to breakdown a KC into multiple ones. While the categorization and viewing of learning curves helps to indicate there is a potential problem, it may not be clear to the user how to optimally resolve the problem.

Acknowledgment. The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

References

1. Baker, R.S., Inventado, P.S.: Educational data mining and learning analytics. In: Learning Analytics, pp. 61–75. Springer, New York (2014)
2. Murray, T.: An overview of intelligent tutoring system authoring tools: updated analysis of the state of the art. In: Authoring Tools for Advanced Technology Learning Environments, pp. 491–544. Springer, Dordrecht (2003)
3. Fancsali, S.E., Ritter, S., Stamper, J., Nixon, T.: Toward "hyperpersonalized" cognitive tutors. In: AIED 2013 Workshops Proceedings Volume, vol. 7, pp. 71–79 (2013)
4. Stamper, J., Koedinger, K.R.: Human-machine student model discovery and improvement using DataShop. In: Kay, J., Bull, S., Biswas, G. (eds.) Proceeding of the 15th International Conference on Artificial Intelligence in Education (AIED 2011), pp. 353–360. Springer, Berlin (2011)

5. Stamper, J., Koedinger, K., Baker, R.S., Skogsholm, A., Leber, B., Rankin, J., Demi, S.: PSLC DataShop: a data analysis service for the learning science community. In: International Conference on Intelligent Tutoring Systems, p. 455. Springer, Heidelberg (2010)
6. VanLehn, K.: The behavior of tutoring systems. Int. J. AIED **16**, 227–265 (2006)
7. Shepard, L.A.: What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In: Gifford, B.R., O'Connor, M.C. (eds.) Changing Assessment: Alternative Views of Aptitude, Achievement and Instruction, pp. 301–328. Kluwer, Boston (1992). 10, 978-94
8. Baker, R.S., Corbett, A.T., Aleven, V.: More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. In: Intelligent Tutoring Systems, pp. 406–415. Springer, Heidelberg, June 2008
9. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: lessons learned. J. Learn. Sci. **4**(2), 167–207 (1995)
10. Cen, H., Koedinger, K.R., Junker, B.: Is over practice necessary? Improving learning efficiency with the cognitive tutor. In: Proceedings of the 13th International Conference on Artificial Intelligence and Education (2007)
11. Koedinger, K.R., Mathan, S.: Distinguishing qualitatively different kinds of learning using log files and learning curves. In: ITS 2004 Log Analysis Workshop, pp. 39–46 (2004)
12. Koedinger, K.R., McLaughlin, E.A., Jia, J.Z., Bier, N.L.: Is the doer effect a causal relationship?: how can we tell and why it's important. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, pp. 388–397. ACM, April 2016
13. Anderson, J.R., Conrad, F.G., Corbett, A.T.: Skill acquisition and the LISP tutor. Cogn. Sci. **13**(4), 467–505 (1989)
14. Cen, H., et al.: Learning factors analysis – a general method for cognitive model evaluation and improvement. In: ITS 2006, pp. 164–175 (2006)
15. Moore, S., Stamper, J., Soniya, G.: Human-centered data science for educational technology improvement using crowd workers. In: Companion Proceedings 9th International Conference on Learning Analytics & Knowledge, pp. 341–347, March 2019
16. Draney, K.L., Pirolli, P., Wilson, M.: A measurement model for a complex cognitive skill. Cogn. Diagn. Assess., 103–125 (1995)
17. Koedinger, K., Stamper, J., McLaughlin, E.: Using data-driven discovery of better student models to improve student learning. In: Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED 2013) (2013)
18. Koedinger, K.R., McLaughlin, E.A., Stamper, J.C.: Automated Student Model Improvement. Int. Educ. Data Mining Soc. (2012)
19. Nguyen, H., Wang, Y., Stamper, J., McLaren, B.M.: Using knowledge component modeling to increase domain understanding in a digital learning game. In: Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019), pp. 139–148 (2019)
20. Stamper, J., Koedinger, K., Mclaughlin, E.: A comparison of model selection metrics in datashop. In: Educational Data Mining, July 2013
21. Koedinger, K.R., Baker, R.S.J.D., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: the PSLC DataShop. In: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.D. (eds.) Handbook of Educational Data Mining. CRC Press, Boca Raton (2011)
22. Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: the PSLC DataShop. In: Handbook of Educational Data Mining, vol. 43 (2010)

23. Murray, R.C., Ritter, S., Nixon, T., Schwiebert, R., Hausmann, R.G., Towle, B., Vuong, A.: Revealing the learning in learning curves. In: International Conference on Artificial Intelligence in Education, pp. 473–482. Springer, Heidelberg, July 2013
24. Rohrer, D.: Interleaving helps students distinguish among similar concepts. Educat. Psychol. Rev. **24**(3), 355–367 (2012)