DataShopping for Performance Predictions

Michael Collins[1(✉)], Kevin A. Gluck[2], and Tiffany S. Jastrzembski[2]

[1] ORISE at Air Force Research Laboratory, 2620 Q Street, Building, 852,
Wright-Patterson Air Force Base, OH 45433, USA
michael.collins.74.ctr@us.af.mil

[2] Air Force Research Laboratory's Cognitive Models and Agents Branch, 2620
Q Street, Building, 852, Wright-Patterson Air Force Base, OH 45433, USA
{kevin.gluck, tiffany.jastrzemski}@us.af.mil

Abstract. Mathematical models of learning have been created to capitalize on the regularities that are seen when individuals acquire new skills, which could be useful if implemented in learning management systems. One such mathematical model is the Predictive Performance Equation (PPE). It is the intent that PPE will be used to predict the performance of individuals to inform real-world education and training decisions. However, in order to improve mathematical models of learning, data from multiple samples are needed. Online data repositories, such as Carnegie Mellon University's DataShop, provide data from multiple studies at fine levels of granularity. In this paper, we describe results from a set of analyses ranging across levels of granularity in order to assess the predictive validity of PPE in educational contexts available in the repository.

Keywords: Performance prediction · Datashop · Repository · Learning optimization · Mathematical models

1 Introduction

In many real world domains, there is a recurring need to educate both workers and students in order to equip them with new skills or to ensure that a baseline standard of knowledge and performance is met. Time and cost are relevant factors when deciding whether continuing education or refresher training should be required, as are the potential risks associated with ignorance or decreased proficiency. This complex and consequential trade space is motivating the development of learning management systems that attempt to improve either the quality or rate at which individuals learn or can allow training to be tailored to specific individuals. One way that these goals can be accomplished is for these systems to employ different mathematical models of learning that can be used to generate predictions of future performance of either an aggregate sample or specific individuals in order to inform education and training decisions. This presents an opportunity for basic cognitive science research to find real-world application.

Psychological research has long noted that humans exhibit certain mathematical regularities when learning new knowledge and skills [2, 8]. The Predictive

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2015, LNAI 9183, pp. 12–23, 2015.
DOI: 10.1007/978-3-319-20816-9_2

Performance Equation (PPE) (described in the following section) is a mathematical model of learning, forgetting, and the spacing effect that uses the historical performance of individuals to generate a prediction of their future performance [4]. The PPE was developed from Anderson and Schunn's [1] General Performance Equation (GPE), which captured the important influences of the power laws of learning and forgetting as determinants of performance. To improve upon the GPE, Jastrzembski et al. [4] introduced the PPE, taking into account the effects on performance of temporal spacing between instances of practice [2]. The PPE has been validated across many samples of data from laboratory studies on learning and retention [5].

A typical limitation of archival publications in the literature regarding learning and retention is that data are only reported at the sample level of analysis. In other words, data regarding measures of central tendency and variability are only reported at the level of the entire sample, or at best the level of the experimentally manipulated sub-samples of interest. This is understandable and generally serves the proximal scientific objective of each particular study quite well. However, most of the real world applications of technologies like learning management systems involve assessment at a finer level of granularity, down at the level of the individual learners. For that level of analysis we generally need source data, rather than archival, summative publications. This is the sort of niche intended to be filled by online data repositories. Repositories can be used to test model predictions and explore new uses of different models of learning. One such data repository is DataShop located on Learnlab.org, created by the National Science Foundation-funded Pittsburgh Science of Learning Center [7]. It holds a large set of publicly available data that can be used to further learning research objectives, and this paper describes a case study using DataShop for exactly that purpose. In this case, we used data available on DataShop to advance our performance prediction research using fine-grained third-party data available on its public database.

A primary intended application of the PPE is in adaptive scheduling of continuing education and refresher training. To accomplish this we will use a real-time calibration update method, which has PPE make multiple sequential predictions about the performance of a sample or individual over time, re-calibrating and updating it prediction of future performance each time a new data point becomes available. Implementing the PPE in this way may be beneficial in two ways. The first possible benefit is, that the PPE can take into account the entire historical performance of a group or individual in order to generate a prediction of what their future performance might be. Using the entire available historical performance of a sample or individual to generate a prediction is useful especially within applied domains where more variability is seen in human performance compared to the performance seen in controlled short-term laboratory studies. The second possible benefit is, by allowing the model to calibrate to all of the available instances of learning, the PPE can make a more informed prediction of what the sample or individual's performance will be during the next event, a prediction that would be useful to inform training decisions. However, it is currently unknown whether the calibration update method gives rise to better predictions over time, or if it allows for informative predictions to be made at different levels of aggregation (e.g., the aggregate performance of an individual across multiple tasks or the aggregate performance of a sample over multiple or single tasks).

In this paper, data collected from thirteen different classroom tutoring studies[1] exported from DataShop were used to perform a validation analysis of PPE using the calibration update method. The PPE was used to predict the performance of multiple samples across several levels of aggregation over different numbers of events, mimicking how PPE would be used in an applied domain. We examined how both the level of aggregation and the number of calibration and prediction events affected PPE's general ability to both fit the initial performance and predict future performance when implemented using the calibration update method.

2 The Predictive Performance Equation

The PPE is a hybrid model of learning and forgetting that predicts future performance by exploiting the mathematical regularities seen when individuals acquire a skill or learn new information, based on the amount of experience they have had on a given task and the amount of time in between instances of practice. The PPE works by calibrating to historical performance data using three free parameters (Eq. 1).

$$\text{Performance} = S * ST * N^{c} * T^{-d} \tag{1}$$

The three free model parameters are S (scalar), used to accommodate the performance measure of interest (e.g., Error Rate, Percent Correct, Response Time, etc.), c (learning rate), and d (decay rate). There are also fixed parameters determined by the timing and frequency of events in the protocol, such as T, the amount of true time passed since the onset of training, and N, the discrete number of training events that occurred in the training period. ST (Eq. 2) is the stability term that "captures the effects of spacing, by calibrating experience amassed as a function of temporal training distribution and true time passed" [6, p. 110].

$$St = \frac{\sum lag}{P} \cdot \frac{P_i}{T_i} \cdot \frac{\sum_i^j (lag_{max_{i,j}} - lag_{min_{i,j}})}{N_i} \tag{2}$$

Lag denotes the amount of wall time that has passed since the last training event and P is the amount of true time amassed during practice. Once the PPE has calibrated to the set of historical data points, it uses the learning and decay rates, the amount of experience on a task(s), and the time since its previous instance of practice to generate a quantitative point prediction of future performance for the sample or individual [5].

[1] The thirteen studies used in this paper are listed in the acknowledgement section, per instructions of DataShop.

3 Studies Collected from DataShop

DataShop is a repository of data collected from different learning and tutoring studies. A single sample on DataShop is referred to as a dataset, which is composed of a record of performance of individuals who attempted to solve a set of problems in a specific domain within a certain period of time [3]. For example, one dataset in the repository contains a record of students' performance when solving physics problems in a tutoring system used at the United States Naval Academy over the Fall 2007 semester. Each dataset contains a record of the performance of each individual across that curriculum's content. Each curriculum is made up of *problems*, defined as "a task [attempted by] a student usually involving several steps" [3]. An example of a problem in the physics tutor is comparing the difference in velocity between train *A* and *B*. Successfully solving a problem involves completing a series of *steps*, which are "an observable part of a solution to a problem" [3], such as finding the velocity of train *A*, which subjects attempted to solve over the course of the study. In the analysis presented here, we examined performance on the individual steps within a dataset, because they represented the distinct pieces of knowledge and skills learners were acquiring, over the course of the study.

3.1 Data Organization

All of the data from each dataset was organized into three different levels of aggregation. The first and highest level of aggregation was a sample's performance across multiple steps, defined as a sample of data from a single dataset composed of individuals who all had the same number of opportunities to complete the same steps. The second level of aggregation was a sample's performance on a specific step, defined as a sample of individuals from the same dataset who all had the same number of opportunities to attempt the same step. The third level of aggregation was an individual's performance across multiple steps, defined as a sample of steps done by the same participant.

Due to the fact that a majority of the datasets used for these analyses were recorded on tutoring systems that students used to complete their homework throughout the semester or year, each individual within a dataset did not have the same number of opportunities to complete each step (e.g., One individual could have had four opportunities to attempt a single step, while another individual had eight opportunities to attempt the same step). Each student did not have the same number of opportunities to attempt all steps because either an individual dropped out of the class or showed a high enough competence on a particular problem and was not presented with any more opportunities to solve that step. In order to create equivalent samples (subsets), that included the same individuals, attempting the same steps, for the same number of times for each of the three levels of aggregation, multiple subsets from each dataset were created. Subsets were constructed based on the number of opportunities that individual participants had with specific steps within a dataset.

For example, one participant could have had six opportunities to attempt steps A and B and eight opportunities with steps C and D, while only having four opportunities to attempt step E. In order to formulate a subset whose performance is aggregated over this individual's data where each step was done for the same number of times, one subset can be composed from this individual's first four opportunities with steps A, B, C, D, and E, creating four unique events where each step was done once. A second subset can also be composed from the individual's first six opportunities with steps A, B, C, and D, creating a subset of data with six unique events. This process of separating the data in order to create multiple subsets comprised of four to nine events was repeated across all thirteen datasets for each level of aggregation.

Error rate was used as a dependent measure, determined by the individual's first attempt at each step (i.e., correct or incorrect). The error rate was calculated by the percentage of incorrect first attempts across all of the steps done at each event within a subset. It is expected that the error rate will be highest at the subset's first event, because it is the first time participant(s) will have attempted to solve these steps. Additionally, error rate should decrease in subsequent events as individual(s) gain more experience solving these steps over time.

After each dataset was sorted into multiple subsets for each of the three different levels of aggregation, two different criteria were used to decide whether a subset was included or excluded from the exploratory analysis. The first criterion was that the subset's performance had to improve over the period of time that it was observed, and was assessed by comparing the error rate of the first and last event. Subsets of data where the error rate was higher on their final event than on their first event were excluded from our analysis, due to the fact that the PPE would not be able to account for their performance. Of all the subsets that were compiled from the thirteen datasets, 18 (26 %) of subsets aggregated over a multiple steps, 1745 (56 %) of subsets aggregated over a single step, and 2006 (47 %) subsets aggregated over an individual were excluded from our analysis for not meeting the first criterion. The second criterion was that the number of steps present at each trial had to be greater than or equal to the number of events within a subset. Thus, if a subset whose performance was aggregated over an individual participant was observed over nine events, their performance had to be aggregated over a minimum of nine individual steps. The same criterion was used for subsets aggregated over a single step, though to be included in our analysis a minimum number of participants were required. A minimum number of steps or participants were required for each subset because it was found that the error rates of subsets aggregated over a few steps or participants (e.g., one or two steps) were highly variable. Minimal changes in performance of subsets aggregated over a few steps or individuals caused large fluctuations in their error rates (e.g., a subsets error rate could go from 100 % on event one to 0 % on event two and 50 % on event three), which PPE could not account for. From the subsets that met the first criterion, 551 (58 %) of subsets aggregated over multiple subsets and 652 (33 %) of subsets aggregated over an individual were excluded for not also meeting the second criterion. All remaining subsets, which met both criteria, were used in the exploratory analysis (Table 1).

Table 1. Shows the total number of subsets comprised of different numbers of events across all three levels of aggregation that were included in the exploratory analysis.

Level of aggregation	Number of events in a subset						Total
	4	5	6	7	8	9	
Subsets aggregated over multiple steps	14	12	12	10	8	10	66
Subsets aggregated over a single steps	331	136	105	81	61	57	771
Subsets aggregated over an individual	410	314	234	203	104	90	1355

3.2 Calibration Updating with the Predictive Performance Equation

The PPE was run the same way on each subset across all three levels of aggregation, following the calibration update method. The PPE began by calibrating to the performance of the first three events of each subset (the minimum number of events needed for the PPE to make an initial prediction), then made a prediction about the subset's performance on their remaining events. After its initial prediction, PPE then updated its prediction by calibrating to all of the previous events plus one additional event and again generated a prediction for the subset's performance on the remaining events. The PPE continued to predict the performance of a subset's remaining events until it had calibrated up to the second to last event.

4 Results

The goal of this study was to examine PPE's ability to both account for and predict performance of subsets across multiple datasets at different levels of aggregation. In order to address these goals, we assessed PPE's ability, (a) to calibrate to the initial performance of a subset (calibration period) and (b) to predict the performance of future events (prediction period). To examine the overall ability of PPE to calibrate to the subsets' initial performance, the R^2 and RMSD of the calibration period for each prediction was computed and averaged over subsets where PPE had calibrated to the same number of events from the same level of aggregation (e.g., subsets whose performance was aggregated over a single individual and where PPE had only calibrated to the performance of the first three events). In order to examine PPE's ability to predict the performance of subsets' future events, the R^2 and RMSD of the prediction portion for each prediction made by PPE was computed and averaged over subsets from the same level of aggregation where PPE had both calibrated to and predicted the performance of the same number of events (e.g., subsets whose performance was aggregated over multiple steps, where PPE had calibrated to its first three events and predicted the performance of on its last three events). The overall R^2 for each of PPE's predictions were not recorded when predicting the performance of a subsets' last two events, because the R^2 of the prediction portion was always one between the PPE's prediction and the subset's performance of their last two events – it is simply a best fitting straight line. The overall R^2 of the prediction period could not be computed when only predicting the performance of the last event because a R^2 could not be calculated with only one event. To examine the change in PPE's predictions of performance of a single event while calibrating to additional instances of performance, the residual between PPE's prediction and a subset's performance on the last event was recorded after each prediction. The absolute value of the last residual of each prediction was then averaged across subsets from the same level of aggregation where PPE had predicted over the same number of events.

4.1 Calibration Period

Calibration period R^2 and RMSD varied only slightly as a function of aggregation. Looking across the three levels of aggregation, the average R^2 for subsets whose performance was aggregated over multiple steps (*M* = .83, *SD* = .20) did not differ greatly from subsets aggregated over a single step (*M* = .78, *SD* = .29) (Fig. 1). The average R^2 of subsets whose performance was aggregated over a single individual (*M* = .73, *SD* = .28) was slightly lower still (Fig. 1). The level of aggregation was found to have a similar effect on the overall RMSD values, with fit quality (in this case meaning a *lower* RMSD) positively correlated with the level of aggregation (Fig. 2). The average RMSD when PPE calibrated to the first three events of a subset's performance was lowest in subsets' aggregated over multiple steps (*M* = .01, *SD* = .03), with little difference being seen in subsets' aggregated over a single step (*M* = .04, *SD* = .06) or individual (*M* = .05, *SD* = .03) (Fig. 2). The results from the average R^2 and RMSD of the calibration period across each of the three levels of aggregation show that the ability of PPE to fit the initial performance of a subset's performance in part depends on the granularity of the data.

The pattern holds across levels of aggregation, as we increase the number of calibration points, but fits deteriorate as the number of calibration data points increases. The highest average R^2 and the lowest RMSD at each level of aggregation, occurred when PPE calibrated only to the performance of a subset's first three events (Figs. 1 and 2). The average R^2 of the calibration period decreased and the average RMSD increased, meaning that the calibration fits get worse, as PPE calibrated to each additional event. The impact of adding calibration events is stronger on the R^2 than on the RMSD, which increased but only slightly as PPE calibrated to additional events. Although, the overall ability of PPE to calibrate to additional events decreased as the number of events calibrated to increased, it is how these results affect PPE's ability to predict future performance that is the more important question when assessing the effectiveness of the calibration update method.

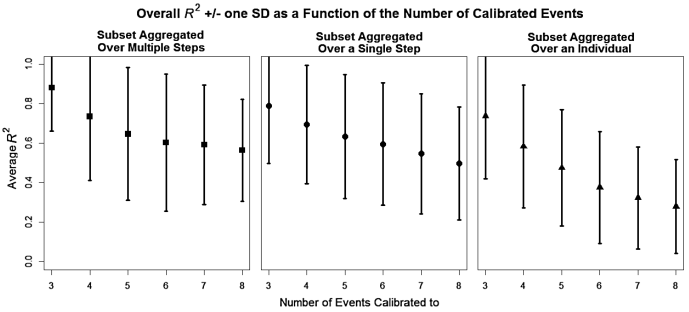

Fig. 1. Shows the average R^2, ± one standard deviation (SD) of the calibration period as a function of the number of events calibrated to, across the three levels of aggregation.

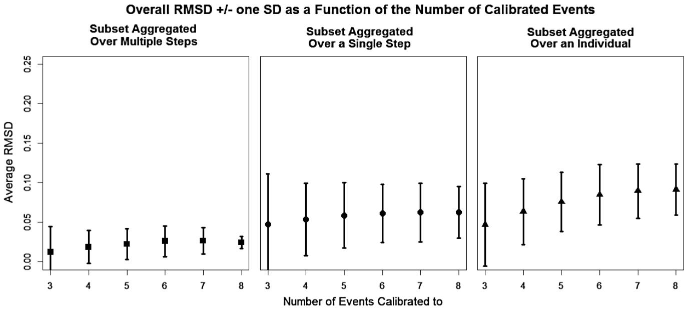

Fig. 2. Shows the Average RMSD, ± one standard deviation (SD) of the calibration period as a function of the number of events calibrated to, across the three levels of aggregation.

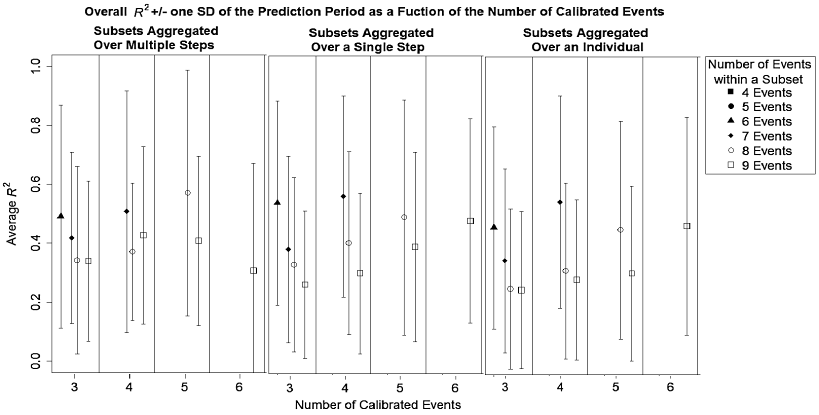

Fig. 3. Shows the average R^2, ± one standard deviation (SD) of the prediction period as a function of the number of events calibrated to, across the three levels of aggregation.

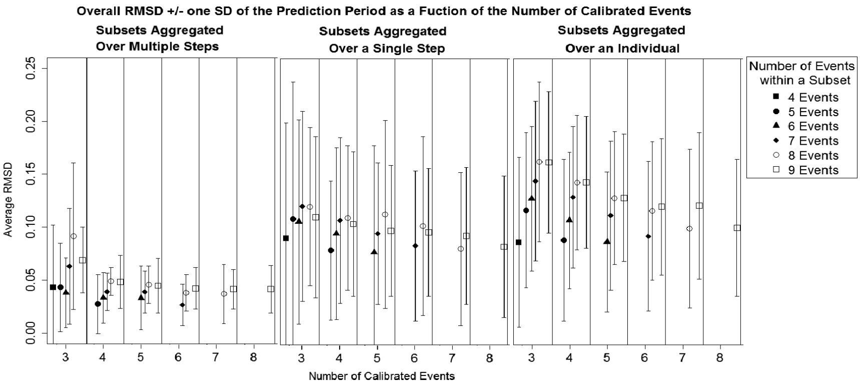

Fig. 4. Shows the average RMSD, ± one standard deviation (SD) of the prediction period as a function of the number of events calibrated to, across the three levels of aggregation.

4.2 Prediction Period

The number of events PPE calibrated to affected its prediction accuracy (Figs. 3 and 4). The worst prediction accuracies (lowest average R^2 values and the highest average RMSD) across all three levels of aggregation occurred when PPE calibrated to the subsets' performance on their first three events and predicted the performance of their remaining events. The R^2 increased, and the RMSD decreased as PPE calibrated to additional events across all three levels of aggregation. The only exception to this trend occurred in subsets aggregated across multiple steps comprised of nine events. In these subsets, the overall R^2 of the prediction period decreased after PPE calibrated to the first five and six events and predicted the performance on the remaining four and three events.

The average last residual between the PPE's prediction of a subset's performance on their final event varied as a function of the level of aggregation (Fig. 5). The smallest average residual across each level of aggregation was found to be when predicting the performance of subsets aggregated over multiple steps (*M* = .04, *SD* = .04), followed by subsets aggregated over a single step (*M* = .08, *SD* = .09) and then subsets aggregated over an individual (*M* = .09, *SD* = .07). The number of events that PPE predicted over was also found to affect the average residual, across all three levels of aggregation, finding that its most accurate predictions occurred when it had calibrated up to the second to last event predicting the performance on a subset's final event and its accuracy decreased as the number of events PPE predicted over increased. Results show that when calibrating to all available historical events in a subset, PPE can generate a more accurate prediction of future performance, though its accuracy in part depends on the level of aggregation of the data it is predicting over.

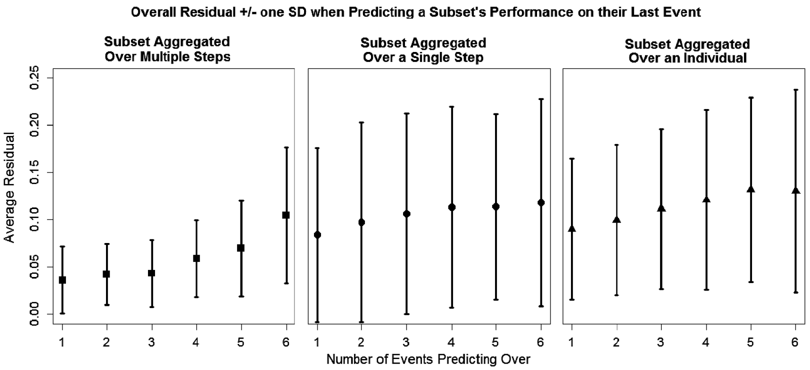

Fig. 5. The average absolute value of the residual between the PPE's predictions and a subset's performance on their last event as a function of the number of events predicting over, across all three levels of aggregation.

5 Discussion

The goal of this study was to perform an exploratory analysis on datasets available on DataShop to implement PPE using the calibration update method across different levels of aggregation, examining its ability to both fit and predict the performance of different subsets. During the calibration period, the level of aggregation and the number of calibration events both affected PPE's overall fit to a subset's initial performance. The PPE calibrated to subsets aggregated over multiple or single steps better than subsets which were aggregated over of an individual, as measured by the average R^2 and RMSD values of the calibration period. The effect the level of aggregation had on the calibration period is understandable due to the noise within a sample being decreased when a subset's performance is averaged over multiple individuals and steps; it is the statistical smoothing benefit of the law of large numbers.

The number of events to which PPE calibrated also affected its ability to fit the initial performance of a subset, decreasing the R^2 and increasing the RMSD with each additional event calibrated to. When PPE calibrates to the initial performance of a subset, it attempts to minimize the RMSD between the subset's initial performance and PPE's calibration, by manipulating its three free parameters. As PPE repeatedly calibrates to additional events within a subset, it must attempt to minimize the RMSD between more events, which may not all clearly fall along a best fit line. This pattern of results was seen across each level of aggregation showing that as PPE calibrates to additional events, it loses some of its ability to fit the initial performance of a subset.

Although as PPE calibrated to additional events its ability to fit a subset's performance decreased, the opposite results were seen during the prediction portion. As the number of events PPE calibrated to increased, so did its overall ability to predict the performance of future events across all three levels of aggregation. The only exception to this trend was observed in subsets aggregated over multiple steps comprised of nine events. This group of subsets was comprised of a higher proportion of steps which learners showed no improvement on, over the period of time they were observed, which lead to inconsistent and overall little improvement in their aggregate performance. The combination of inconsistent improvement and a small overall learning that was seen in the performance of these subsets affected the R^2 of the prediction portion. However, the RMSD of the prediction portion of these subsets did not increase as PPE continued to calibrate to additional events, suggesting that PPE was not able to predict the variability in their performance, though it was able to capture the overall trend in the performance of these subsets.

The accuracy of PPE's predictions of performance on a single event were also found to depend on the subset's level of aggregation. The PPE's most accurate predictions were made when predicting subsets aggregated over multiple steps; its accuracy decreased when predicting subsets aggregated over a single step or individual. Across all three levels of aggregation, predictions of performance of a single event were most accurate when PPE calibrated to all but one of the events within a subset and predicted the performance on its final event and its accuracy decreased as the number of events it predicted over increased. The results from both the accuracy of PPE predictions of a single event, the R^2, and RMSD of the prediction period suggest that by calibrating to additional events, PPE can better predict the future performance of a subset, despite the decrease in its ability to calibrate to the additional instances of a subset's initial performance.

The results obtained from using PPE on the datasets collected from DataShop shed light on two points for our prediction performance research. The first was, across each level of aggregation, PPE's predictions were overall improved by continually calibrating to additional events, in order to update its predictions. The second was, that the accuracy of PPE's predictions were able to be examined across three different levels of aggregation, which were seen to affect the accuracy of its predictions. If mathematical models of learning are to be implemented in learning management systems developed to help improve the rate or quality of training and education that specific individuals receive, it is not enough for them to be able account for a sample's aggregate performance, but the performance at lower levels of aggregation, such as the performance of an individual on a single task or the individuals within a sample, need to be able to be accounted for as well. One way to account for the inherent variability in the performance seen in data at low levels of granularity is to include prediction intervals in addition to making quantitative point predictions of future performance. Prediction intervals would allow PPE to predict a range that future performance might fall within, depending on both the granularity of the data and time period predicting over, two factors which were seen to affect PPE's predictions. Jastrzembski et al. [6] have previously addressed the need for incorporating prediction intervals into PPE. The results reported here show the additional benefit that prediction intervals could add, by being able to address the uncertainty seen in the performance at low levels of aggregation.

In conclusion, the data collected from DataShop allowed for a large scale exploratory analysis to examine the utility of implementing PPE using the calibration update method. Datasets with a record of each individuals' performance at the individual-step level allowed for a far more in depth analysis of PPE's predictions, which archival aggregate sample data from the published literature would not have allowed for. We hope that others will use this case study as an example of how DataShop's public database can be used as a source of data from applied domains and will take advantage of the data being available at a such a fine level of granularity to improve other mathematical models of learning, so that they may find real world application.

Acknowledgements. We would like to thank Carnegie Learning, Inc., for providing the Cognitive Tutoring data supporting this analysis. We used the “Handwriting/Examples Dec 2006”, “Geometry Area (1996–1997)”, “USNA Physics Fall 2006”, “USNA Physics Fall 2007”, “USNA Introductory Physics Fall 2009”, “USNA Physics Spring 2007”, “USNA Physics Spring 2008”, “USNA Introductory Physics Spring 2010”, “Watchung Hills Regional High School AP Physics 2007–2008”, “Watchung Hills Regional High School AP Physics 2008–2009”, “Watchung Hills Regional High School Honors Physics 2007–2008”, “Watchung Hills Regional High School Honors Physics 2008–2009” and “Watchung Hills Regional High School Honors Physics 2009–2010” accessed via DataShop. We would also like to thank Michael Krusmark for the work he was done implementing the PPE into R. This research was supported by the Air Force Research Laboratory through AFOSR grant 13RH15COR and through Early Career Award Winner funds provided to the third author.

References

1. Anderson, J.R., Schunn, C.: Implications of the ACT-R learning theory: no magic bullets. In: Glaser, R. (ed.) Advances in Instructional Psychology: Educational Design and Cognitive Science, vol. 5, pp. 1–33. Erlbaum, Mahwah (2000)
2. Bahrick, H.P., Bahrick, L.E., Bahrick, A.S., Bahrick, P.E.: Maintenance of foreign language vocabulary and the spacing effect. Psychol. Sci. **4**, 316–321 (1993)
3. Glossary. https://pslcdatashop.web.cmu.edu/help?page=terms
4. Jastrzembski, T.S., Gluck, K.A., Gunzelmann, G.: Knowledge tracing and prediction of future trainee performance. In: 40th I/ITSEC Annual Meeting, pp. 1498–1508. National Training Systems Association, Orlando, FL (2006)
5. Jastrzembski, T.S., Gluck, K.A.: A formal comparison of model variants for performance prediction. In: Proceedings of the International Conference on Cognitive Modeling (ICCM), Manchester, England (2009)
6. Jastrzembski, T.S., Addis, K., Krusmark, M., Gluck, K.A., Rodgers, S.: Prediction intervals for performance prediction. In: International Conference on Cognitive Modeling (ICCM), pp. 109–114, Philadelphia, PA (2010)
7. 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, pp. 43–56. CRC Press, Boca Raton (2010)
8. Newell, A., Rosenbloom, P.S.: Mechanisms of skill acquisition and the law of practice. In: Anderson, J.R. (ed.) Cognitive Skills and Their Acquisition, vol. 1, pp. 1–55. Lawrence Erlbaum Associates, Inc., Hillsdale (1981)