Artificial Intelligence in Education
C.-K. Looi et al. (Eds.)
IOS Press, 2005

On Using Learning Curves to Evaluate ITS

Brent Martin[1], Kenneth R Koedinger[2], Antonija Mitrovic[1] and Santosh Mathan[2]

[1]*Intelligent Computer Tutoring Group, University of Canterbury,*
Private Bag 4800, Christchurch, New Zealand
{brent,tanja}@cosc.canterbury.ac.nz

[2]*HCI Institute, Carnegie Mellon University, Pittsburgh, PA 15213*

Abstract. Measuring the efficacy of ITS can be hard because there are many confounding factors: short, well-isolated studies suffer from insufficient interaction with the system, while longer studies may be affected by the students' other learning activities. Coarse measurements such as pre- and post-testing are often inconclusive. Learning curves are an alternative tool: slope and fit of learning curves show the rate at which the student learns, and reveal how well the system model fits what the student is learning. The downside is that they are extremely sensitive to changes in the system's setup, which arguably makes them useless for comparing different tutors. We describe these problems in detail and our experiences with them. We also suggest some other ways of using learning curves that may be more useful for making such comparisons.

1 Introduction

Analysing adaptive educational systems such as Intelligent Tutoring Systems (ITS) is hard because the students' interaction with the system is but one small facet of their education experience. Pre- and post-test comparisons provide a rigorous means of comparing two systems, but they require large numbers of students and a sufficiently long learning period. The latter confounds the results unless it can be guaranteed that the students do not undertake any relevant learning outside the system being measured. Further, such experiments can only make comparisons at a high level: when fine-tuning parts of an educational system (such as the domain model), a large number of studies may need to be performed. In our research we have explored using a more objective measure of domain model performance, namely learning curves, to see if we can predict what changes could be made to improve student performance, including at the level of individual rules, or sets of rules. This often involves comparing disparate systems. In particular, we are interested in methods for comparing systems that work for small, short studies, so that we can propose, implement, test and refine improvements to our systems as rapidly as possible to make them maximally effective. The use of learning curves appears attractive in this regard.

Researchers use numerous methods to try to evaluate educational systems. Pre- and post-testing is commonly tried, but the results are often inconclusive. Often other differences are found in how students interacted with the system, but they appear to have been too little to give a clear test outcome. Ainsworth [1] failed to find significant pre-/post-test differences between REDEEM and CBT, but did find differences in certain situations. Similarly, Uresti and duBoulay [8] use pre-/post-testing to determine the efficacy of their learner companion across a variety of variables. They find no significant difference in learning outcome, but do find differences in measurements of usage within the tool.

Suraweera and Mitrovic [7] found significant differences between using their ITS (KERMIT) versus no tutor.

Because of the lack of clear results, researchers often measure other aspects of their systems to try to find differences in behaviour. However, these do not always measure learning performance specifically. Uresti and duBoulay measured the amount their "learning companion" was taught by the student during the session, which is arguably (but not explicitly) linked to improved learning. Walker et al [9] performed post-hoc analysis of the predictive ability of their collaborative information filter (which measures how well it chooses material), but they do not measure the effect on learning. Zapata and Greer [10] evaluated their inspectable Bayesian student modelling method by observation of the actions students performed and their interactions with the system, but again this does not measure changes in learning performance. Finally, many studies include the use of questionnaires to analyse student attitudes towards the system.

The use of learning curves attempts to bridge this gap by measuring learning activity within the system. As well as showing how well a particular system supports learning, they have the potential to allow quantitative comparisons between disparate systems. However, there are problems with such comparisons that need to be overcome. It is hoped that a better understanding of these curves and their limitations will add to the range of evaluative tools at our disposal.

Section 2 describes the use of learning curves for measuring ITS performance. We then describe the specific problems with comparing systems in Section 3, and examine some possible solutions, followed by a discussion in Section 4. Finally, we present our conclusions in Section 5.

2 Learning Curves

Learning curves plot the performance of students with respect to some measure of their ability over time. In the case of ITS, the standard approach is to measure the proportion of knowledge elements in the domain model applied by the student that have been used incorrectly, or the "error rate". Alternatives exist, such as the number of attempts taken to correct a particular type of error. Time is generally represented by the number of occasions the knowledge element has been used. This in turn may be determined in a variety of ways: for example, it may represent each *new problem* the student attempted that was relevant to this knowledge element, on the grounds that repeated attempts within a single problem are benefiting from the user having been given feedback about that particular circumstance, hence they may improve from one attempt to the next by simply carrying out the suggestions in the feedback without learning from them. If the student is learning the knowledge elements being measured, the learning curve will follow a so-called "power law of practise" [6]. Evidence of such a curve indicates that the student is learning the knowledge elements, or, conversely, that the elements represent what the student is learning: a poor power law fit suggests a deficient domain model. Therefore, when comparing two models we might argue that the model showing better power law fit is somehow superior.

The formula for a power law is:

$$Y = Ax^{-B} \qquad (1)$$

The constant *A* represents the Y axis intercept, which for learning curves is the error rate at x=1, or the error rate prior to any practise. *B* depicts the power law slope, equivalent to the linear slope when the data is plotted using a log-log axis. This indicates the steepness of the curve, and hence the speed with which the student is learning the material. Finally, the fit of the power law to the data is measured. All of these may be used to compare two different approaches to determine which is better.

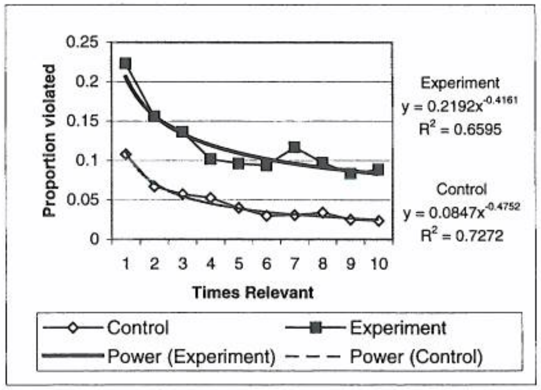

Fig. 1. Learning curves for two variants of SQL-Tutor

Data for learning curves is usually obtained post-hoc from student logs. For each student, a trace is generated for each knowledge element indicating the degree to which the student has correctly applied it. This may be a continuous value or simply "satisfied" or "violated". Data values for a single knowledge element for a single student are unlikely to produce a smooth power law; they simply represent too little data. However, the data can be aggregated in several ways to represent useful summaries: data can be grouped for all students by knowledge element (to compare individual elements for efficacy), by student over all elements (to compare students) or over both for comparing different systems (e.g. two different domain models). The power law fit and slopes can then be compared. Fig. 1. illustrates this: the two curves represent the learning histories for two populations using different variants of the same ITS (SQL-Tutor [5]). The curve has been limited to the first 10 problems for which each constraint is relevant. This is necessary because aggregated learning curves degrade over time because the number of averaged data points decreases. Both curves exhibit a similar degree of fit, and their exponential slopes are similar. However, the Y asymptotes are markedly different, with the experimental group exhibiting more than double the initial error rate of the control group.

3 Problems with Comparing Models

Whilst it appears that learning curves can be compared with one another, there are several issues that call this practise into question. When comparing two different domain models, the power law parameters of fit and slope may be affected by incidental differences that arguably do not affect the quality of the model. These are now explored.

3.1 Fit versus Data Size

The quality of a power law tends to increase with data set size. A larger domain model is therefore likely to exhibit a better fit than a smaller one, even if it does not teach the student any better. For example, Koedinger and Mathan [3] compared learning outcomes associated with two types of feedback in the context of a spreadsheet tutor (an example of a cognitive tutor [2]). In the *Expert* version of the tutor, students were given corrective feedback as soon as they deviated from an efficient solution path. In the *Intelligent Novice* version, students were allowed to make errors; feedback was structured to guide students through error detection and correction activities. A learning curve analysis was performed to determine whether students in one condition acquired knowledge in a form that would generalize more broadly across problems. The tutor provided opportunities to practice six types of problems. A shallow mastery of the domain would result in the acquisition of a unique rule for each type of problem. A deeper understanding of domain principles would allow students to see the common abstract structure in problems that may seem superficially different. Consequently, students would acquire a smaller set of rules that would generalize across multiple problems. In the case of the spreadsheet tutor it was possible to use a set of four rules to solve the six types of problems represented in the tutor.

Two plots were created (Fig. 2), each with a different assumption about the underlying encoding. One plot assumed a unique rule associated with each of the six types of problems represented in the tutor. Thus, with each iteration through the six types of problems, there was a single opportunity to apply each production rule. In contrast, with a four skill, deep encoding, there were multiple opportunities to practice production rules that generalize across problems. Fitting power law curves to data plotted with these alternative assumptions about the underlying skill encoding might determine whether or not students were acquiring a skill encoding that would generalize well across problems.

Both graphs strongly suggest that the "intelligent novice" system is considerably better than the "expert" version – both fit and slope are considerably higher for this variant. However, the difference between the six- and four-skill models is not so clear. For both the expert and novice systems, the slope is higher for the four-skill model, suggesting more learning took place: this is particularly true for the "expert" system. However, in both cases the fit *decreases*, and again this is more marked in the "expert" system. At first glance these observations appear contradictory: learning is improved but quality of the model (as defined by fit) is lower. However, the four-skill model has 33% fewer knowledge elements than the original model, so we would expect the fit to degrade. This means we are unable to make comparisons based on fit in this case. Further, the comparisons of slope now arguably also become dubious. This latter concern could be overcome by plotting individual student curves and testing for a statistically significant difference in the average slopes, as described in Section 3.2.

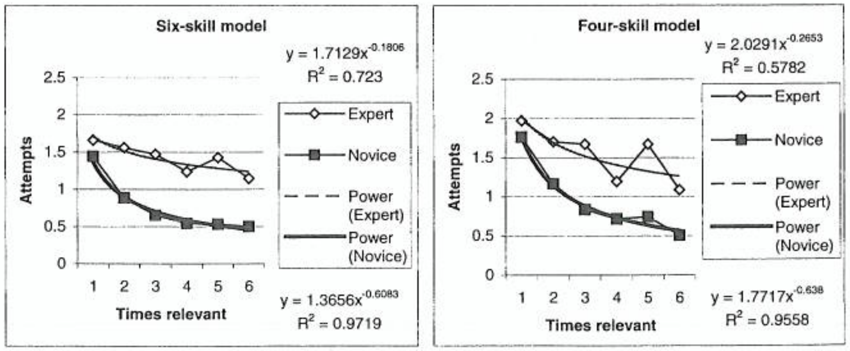

Fig. 2. Learning curves for six- versus four-skill models of the Excel tutor.

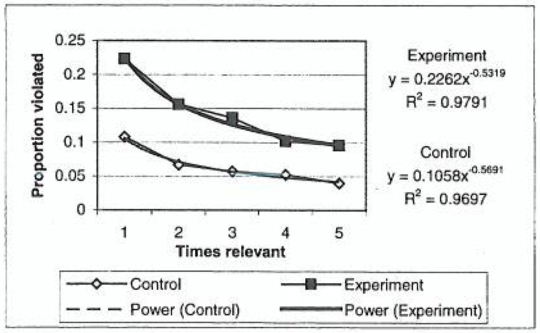

Fig. 3. Two variants of SQL-Tutor with different domain models

3.2 Initial versus Exponential slope

A serious issue with the use of power law slope is that it is highly sensitive to changes in the other parameters of the curve, particularly the Y axis intercept. In [4], we compared two versions of SQL-Tutor that had different problem sets and selection strategies. Fig. 3 shows the learning curves for the two systems trialled on samples of 12 (control) and 14 (experiment) University students. The two curves have similar fit and slope, which might lead us to conclude there is little difference in performance. However, the raw reduction in error suggests otherwise: between x=1 and x=5, the experimental group have reduced their error rate by 0.12, whereas the control group has only improved by 0.7, or about half.

The problem is that power law slope is affected by scale. Fig. 4 illustrates what happens if we modify the scale of a curve by multiplying each data point by two. Although this now represents twice the error reduction over time, the exponential slope is virtually unchanged. Further, adding a constant to the same data *reduces* the exponential slope considerably, even though the net learning is the same. In the case of our study, we were measuring differences caused by an improved problem selection strategy: if the new strategy is better, it should cause the student to learn a greater volume of new concepts at a time. The power law slope does not measure this. However, the Y axis intercept *does* reflect this difference, because it measures the size of the initial error rate. We argued therefore that by comparing the slope of the curve at x=1, we are measuring the reduction in error at the beginning of the curve, which represents how much the student is learning in absolute terms. For the graphs

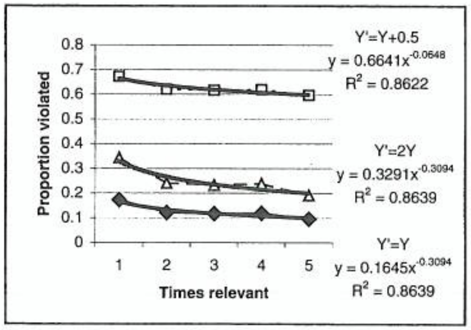

Fig. 4. Scale effects on learning curve slope

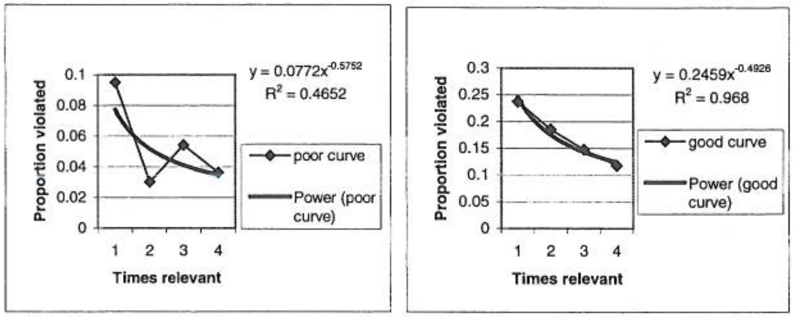

Fig. 5. Examples of individual student learning curves

in Fig. 4 this gives initial slopes of 0.12 for the experimental group and 0.06 for the control group, which correlates with the overall gain for x=5. The advantage of using initial slope rather than simply calculating the gain directly is that the former is using the best fit curve, which averages out errors across the graph, while the latter is a point calculation and is therefore more sensitive to error.

The fact that we have averaged the results across both all knowledge elements and students (in a sample group) may raise questions about the importance of the result. This is measured by plotting curves for individual students, calculating the learning rates and comparing the means for the two populations using an independent samples T-test. Fig. 5 shows examples of individual student curves. In general the quality of curves is poor because of the low volume of data, although some students exhibit high-quality curves. We have noticed a positive correlation between curve fit and slope. For the experiment described this yielded similar results to the averaged curves (initial learning rate = 0.16 for the experimental group and 0.07 for the control group). Further, the T-test indicated that this result was significant (p<0.01). We can therefore be confident that the experimental group exhibited faster learning of the domain model.

3.3 Early versus absolute learning

When evaluating learning curves, we assume that the power law of practise holds, and that the students' error rate will therefore trend towards zero errors in a negative exponential curve. However, there are arguably *two power laws superimposed*: the first is caused by simple practice, and should eventually trend to zero, although this may take a very long time. The second is caused by the feedback the system is giving: as long as this feedback is effective the student will improve, probably following a power law. However, we do not know how the effect of the feedback will vary with time: if it becomes less effective, the overall curve will "flatten", and thus deviate from a power curve. Even if the effect of feedback is constant (and therefore a curve based on feedback effect but not practice effect would trend to zero,) this curve may trend downwards much faster than the practice curve, and so will eventually intersect, and then be swamped by, the practise curve. The overall graph will therefore appear to be a power law trending to a Y asymptote greater than 0.

Fig. 6 illustrates this point. In this study, we compared two different types of feedback in SQL-Tutor on samples of 23 (control) and 24 (experiment) second year University students. The control system presented the student with the standard (low-level) feedback, while the experimental system grouped several related knowledge elements together, and gave feedback at a more abstract level.

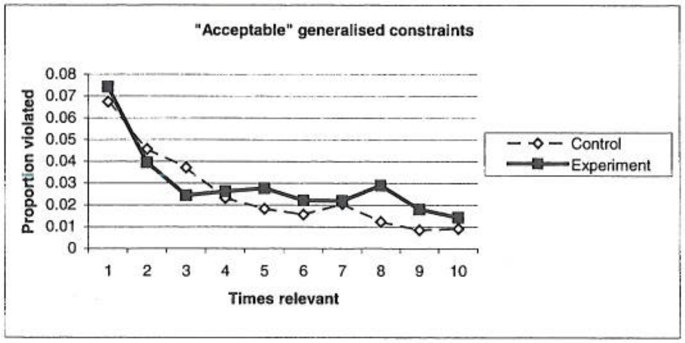

Fig. 6. Comparison domain models with differing feedback granularity

Over the length of the curves the amount of learning appears comparable between the two systems. However, the absolute gain for the *first two times the feedback was given* (i.e. the difference in Y between x=1 and x=3) is different for the two systems: For the control group the gain is around 0.03, while for the experimental group it is 0.05. We also notice that the curve for the experimental group appears to abruptly flatten off after this, suggesting that the feedback is only effective for the first two times it is viewed; after that it no longer helps the student.

We could use the initial learning rate again to measure the early gain, but this is unlikely to be useful because of the way the curve flattens off, and therefore deviates from the initial trend. (We could cut off the curve at x=3 but this is dubious since it is too few data points.) In this case we used the raw improvement as described in the previous paragraph. We obtained learning curves for individual students and performed a T-test on the value of error(t=3)-error(t=1) for each student. The results were similar to those from the aggregated graphs (mean error reduction = 0.058 for the experimental group and 0.035 for the control group), and the difference was significant (p<0.01).

4 Discussion

Section 3 illustrates some of the problems with comparing disparate systems using learning curves. These difficulties can be summarised into two main obstacles. First, changing the knowledge units being measured can affect the learning curves, even if there is no difference in learning. Conversely, learning differences may be masked by incidental effects. Consider, for example, two domain models that are identical, except that one of them includes a large number of trivially satisfied rules. For example, these rules might be useful in a different population, but turn out to be already known by the current students. These will have the effect of reducing the measured error rate, which leads to an *increase* in the exponential slope of the learning curve when compared to the model lacking these concepts, even though there is no improvement in learning. Further, it could be argued that this model is *worse* in the context of the current population. This could be alleviated by measuring the raw number of errors rather than the proportion of applied concepts that were incorrectly used, but such a measure would then depend on the overall size of the two systems being comparable, to say nothing of the number of concepts being applied at any one time. Thus a bias would appear towards more coarse-grained models. What is needed is some sort of normalisation of the curves.

The second problem is that the curves depend on both the domain model and the problems being set, as illustrated in [4]: setting hard problems involving the appropriate concepts appears to lead to steeper curves. To compare two domain models *only* would therefore require that the exact same problems are set, but this raises the spectre of the sequence of questions being better suited to one or other model.

There is also the question of what should be measured. With respect to fig. 6, it could be argued that the early differences in the curves are a detail only, and that overall learning is worse for the experimental group. However, the ideal behaviour of an education system's feedback arguably does *not* follow a power law: in the perfect system, the students would learn all concepts perfectly after seeing the feedback *once*. Further, gains at any point in the curve indicate superior behaviour in a limited context. In our case, the results suggest we should use general feedback the first few times it is presented; if the student still has problems with a concept, we should switch to more specific feedback. This is an important finding that warrants further investigation.

5 Conclusions

We have shown that education systems can be compared by using learning curves to measure the speed with which students learn the underlying domain model. However, if the systems being compared have different domain models, such comparisons are fraught with problems because of scaling effects; some means of normalising the curves is necessary if such comparisons are to be valid. Until this happens they should be presented with caution and treated with some scepticism. However, if the domain model is the same in the two systems, they can be directly compared.

Finally, we have not presented any empirical evidence that effects measured in learning curves translate into real differences in learning. Comparative studies using both learning curves and pre-/post-testing are needed to establish the relationship between learning curves and actual learning performance.

References

[1] Ainsworth, S.E. and Grimshaw, S., *Evaluating the REDEEM Authoring Tool: Can Teachers Create Effective Learning Environments?* International Journal of Artificial Intelligence in Education, 2004. 14(3): p. 279-312.

[2] Anderson, J.R., Corbett, A.T., Koedinger, K.R., and Pelletier, R., *Cognitive Tutors: Lessons Learned.* Journal of the Learning Sciences, 1995. 4(2): p. 167-207.

[3] Koedinger, K.R. and Mathan, S. *Distinguishing qualitatively different kinds of learning using log files and learning curves.* in *ITS 2004 Log Analysis Workshop*. 2004. Maceio, Brazil. p. 39-46.

[4] Martin, B. and Mitrovic, A. *Automatic Problem Generation in Constraint-Based Tutors.* in *Sixth International Conference on Intelligent Tutoring Systems*. 2002. Biarritz: Springer. p. 388-398.

[5] Mitrovic, A. and Ohlsson, S., *Evaluation of a Constraint-Based Tutor for a Database Language.* International Journal of Artificial Intelligence in Education, 1999. 10: p. 238-256.

[6] Newell, A. and Rosenbloom, P.S., *Mechanisms of skill acquisition and the law of practice*, in *Cognitive skills and their acquisition*, J.R. Anderson, Editor. 1981, Lawrence Erlbaum Associates: Hillsdale, NJ. p. 1-56.

[7] Suraweera, P. and Mitrovic, A., *An Intelligent Tutoring System for Entity RelationshipModelling.* International Journal of Artificial Intelligence in Education, 2004. 14(3): p. 375-417.

[8] Uresti, J. and Du Boulay, B., *Expertise, Motivation and Teaching in Learning Companion Systems.* International Journal of Artificial Intelligence in Education, 2004. 14: p. 67-106.

[9] Walker, A., Recker, M., Lawless, K., and Wiley, D., *Collaborative Information Filtering: a review and an educational application.* International Journal of Artificial Intelligence in Education, 2004. 14(1): p. 3-28.

[10] Zapata-Rivera, J.D. and Greer, J.E., *Interacting with Inspectable Bayesian Student Models.* Artificial Intelligence in Education, 2004. 14(2): p. 127-163.