Mining the Dynamics of Student Utility and Strategy Use during Vocabulary Learning

PHILIP I. PAVLIK JR.
Department of Psychology
University of Memphis
Psychology Building, Rm 434
Memphis, TN 38152
(901) 678-2326
Fax: (901) 678-2579
ppavlik@memphis.edu

This paper describes the development of a dynamical systems model of motivation and metacognition during learning, which explains some of the practically and theoretically important relationships among three student-engagement constructs and performance metrics during learning. In order to better calibrate and understand the model, the model was also fit with additional fixed factor predictors determined from the factor scores from a factor analysis of the pre-survey given to students. This work mined data from computerized adaptive flashcard learning system to create the dynamical systems model. This flashcard practice included pop-up survey questions on the student's experience of recent easiness, strategy use, and usefulness, in addition to the correctness performance data for the practice. The dynamical systems model of this data was then used to simulate various student profiles to predict how they would experience the flashcard system. These simulations show how strategy use in this task is crucial because of the ways it influences performance and perceived usefulness. In the model, this result is shown by a bifurcation for higher and lower strategy use, where the higher strategy use equilibrium is accompanied by performance predictions suggesting learning that is more efficient. In addition, we examined the implications of our data for the flow theory of optimal experience by testing models of this theory and comparing it to a Vygotskian perspective on the results.
Key Words: Dynamical system model, language learning, motivation, metacognition, efficacy, and utility

---

1. INTRODUCTION

Many researchers agree that motivation is a dynamic construct, changing from moment to moment in response to the stream of events students experience while learning [e.g. Witherspoon et al. 2007]. As learning events progress, each period marks changes in student ability, attitudes, perceptions and actions with an underlying continuity across time as these states shift not randomly, but as a function of prior ability, attitudes, perceptions and actions. If this is an accurate description, the system in which students learn can be described as a dynamical system [Ward 2002]. Within a dynamical system, motivationally relevant variables play a crucial role, either fostering or impeding performance or the learning processes. Because we can suppose that a dynamical system has an optimum balance, the dynamical systems perspective maps well to Csikszentmihalyi's framing of the state of flow that occurs for people during activities that balance skills and challenges. Flow is defined as a state of optimal experience that occurs when the skill and challenge are balanced appropriately [Csikszentmihalyi 1990].

This paper attempts to develop the discussion of the theory of flow, by introducing a mathematical model that supposes skill, challenge, and other variables feed into the changing experience of motivation and metacognition in a student during language vocabulary learning. The goal of this work is to show how a dynamical system model provides an innovative way to study moment-to-moment change in motivation and metacognitive variables in the vocabulary-learning task, and to compare the results of the study with the verbal theory used by Csikszentmihalyi to support the flow construct.

As noted, flow is often described by Csikszentmihalyi as optimal experience, and while optimal experience may be motivating, it is not clear how flow theory also encapsulates a theory of motivation. A key part of this alignment is to note that flow, as optimal experience, is rewarding by definition. The rewarding aspect of flow and the fact that optimality of experience is a positive consequence means that flow is motivating. Once a student finds ways to achieve flow during an educational activity, they should become more likely to be engaged with that activity in the future. As suggested by this theory, the data below shows that a measure of the utility of the vocabulary practice activity is correlative with further engagement in the practice as measured by self-reported strategy use. Utility is measured, in this case, by simply asking the student for a rating of usefulness of the recent practice on a Likert-type scale. One theoretical claim this paper makes is that these usefulness ratings tap into flow during learning, because flow is optimal experience, and a significant component of optimal experience is positive utility, or more colloquially, usefulness. Prior research tends to support this claim. For example, in a survey of college students, perceived instrumentality, or usefulness, predicted both intrinsic and extrinsic valuing, even when controlling for learning and performance goals [Miller et al. 1999].

To understand how usefulness occurs as a function of task value it is also helpful to consider the similarity of flow to the Vygotskian construct of zone of proximal development (ZPD) [Vygotsky 1978]. Roughly speaking, in both flow and ZPD theories you do not want challenge to exceed skill, or skill to exceed challenge, but rather you want to find an optimal balance, e.g. a range where performance or learning is possible but the task has not yet been mastered. In this balanced range, according to these theories, optimal experience is manifested as intrinsic and extrinsic values are maximally satisfied. While flow theory seems like a better tool to understand our results in this dynamic continuous practice task, as we will see in the results and the discussion, the importance of strategy use that was found, and the fact that some students appeared to have weak

strategy knowledge to apply to this task will lead us to reconsider a Vygotskian explanation, since one interpretation of our results is that for some students the strategic requirements of the task may have made optimal performance impossible because the task was not within some students ZPDs.

Framing the work in this paper in terms of flow is also difficult because of the paucity of experimental results dealing with this construct directly in a controlled fashion. This may be due to the fact that the flow theory, rather than being specified in the socio-cognitive tradition in specific experimental research paradigms, comes from humanistic and personality psychology, and is often presented as a self-help philosophy and as a theory of experience rather than at the level of a mechanistic model of the socio-cognitive system [Csikszentmihalyi 1997]. This breadth (as a theory of experience for example), while suggestive of the power of the construct, becomes a challenge to researchers hoping to engage with theory of flow, since with such a broad scope it is difficult to narrow in on specific claims of the theory. In other words, what is the core of the theory? For example, in Baker, D'Mello, Rodrigo, and Graesser [Baker et al. 2010] they describe a flow like construct called "engaged concentration" and differentiate from flow by specifying that it need not involve task related aspects of flow such as clear goals, balanced challenge, or immediate direct feedback. Further, they exclude various components of flow that reference its intensity, such as time distortion or loss of self-consciousness.

This paper also qualifies the theory of flow, but not so narrowly as to create a new construct such as "engaged concentration", which, while a characteristic of flow, does not interweave a theory of the probable antecedents of this engaged concentration. Rather, the flow described in this paper maps to Csikszentmihalyi's central claim that flow occurs when the balance between skill and challenge leads to optimal experience [Csikszentmihalyi 1990; Csikszentmihalyi 1997]. However, as with Baker, et al. [Baker, D'Mello, Rodrigo and Graesser 2010], it does seem necessary to restrict our discussion to what might be called "ordinary flow" in that the work here does not presuppose the exceptionally intense flow that might be necessary to cause time distortion, merging of action and awareness, or a loss of self-consciousness. On the other hand, the attempted alignment of the research here with flow does include the possibility that aspects of the task like clear goals or feedback may drive the perception of flow to the extent that these task aspects allow the perception of the useful balance between skills and challenge that this paper attempts to investigate.

Because flow is inherently a temporal construct that unfolds over time spent in an activity, this paper resonates well with the general project of dynamical systems research in the social and cognitive sciences [Vallacher and Nowak 2007; Ward 2002]. However, unlike prior work, it is focused on a generalizable model with educational implications. Prior work helps to set a foundation for this paper, but prior social science work with dynamical systems models tends to lack a connection to data that lends itself well to applications [Gottman et al. 2002; Liebovitch et al. 2008]. Indeed, a search of the literature revealed only one instance of an educationally relevant dynamical model that was fit to data [Guastello et al. 1999], in which the authors chose to fit the system to each learner, a procedure that produces a distribution of dynamical systems. In addition, that study had several other differences, making it difficult to compare with the work here, such as not using cross-validation to check models and using different data collection methodology.

2. DATA COLLECTION

The data mined to create the dynamical system in this paper comes from a Chinese vocabulary tutoring system used in a Chinese course. This task was superficially straightforward, because students saw an "optimal" schedule of drill practice (test with a review if incorrect) and had to respond to each drill practice by typing the matching Chinese pinyin phonetic representation or English meaning. Additionally, because the hypothesis was that strategies and motivation might influence this practice, the software asked three questions , which cycled for each user (using individual random orders for each subject, i.e. Q1, Q2, Q3 or Q2, Q1, Q3, etc.…). One of these questions popped up every 2 minutes during the flashcards, and required only a single number key response to minimize student inconvenience. These questions were as follows:

- "How easy was the recent practice?", with Likert like items ranging from "too hard" to "too easy", on a 5 point scale.
- "How useful for learning was the recent practice?" with Likert like items ranging from "not useful" to "very useful", on a 5-point scale.
- "Were you able to use any learning strategies during the recent practice?" with Likert like items ranging from "mostly used repetition" to "mostly used strategies", on a 5 point scale.

These questions were meant to be unthreatening to the student, since they are only admissions about recent practice, and so should not carry strong social norms for responding one way or another.

These questions attempted to measure three crucial theoretical determinants of performance and motivation that have been suggested over the last 50 years in the literature on behavior, motivation, and metacognition. The first question about easiness gives a measure of the level of efficacy the student feels while doing the work. Efficacy for a task is a technical term synonymous with the expected easiness, so it seemed current experience of easiness would be an excellent proxy for this measure. According to Bandura's theory, efficacy is specific to particular tasks or circumstances and this question addresses this construct by asking about recent practice, which should be more accurate than less proximal measures, since recent easiness taps into recent enactive mastery, hypothesized to be an important source of self-efficacy [Bandura 1997; Pajares 1996]. In the context of flow theory, easiness should measure the level of challenge the individual experiences during the practice.

Regarding the second question, the usefulness, utility, or value of a behavior is often supposed to be a sine qua non factor in determining behavior. For instance, the ACT-R computational modeling system [Anderson and Lebiere 1998] is founded on the idea that production rules (e.g. the actions that process the stimuli and trigger recall of the English meaning) are controlled by the utility of each action, and these utilities grow when production success triggers some reward signal [Anderson et al. 2004]. Since the students' goal (source of reward) in the vocabulary task is learning (students are not graded on performance, only participation in practice), one might expect students to rate the utility of their work according to some estimate of their flow during learning and how useful this learning will be to getting better test grades or to future experiences where they will need to know Chinese. For this reason, the question about usefulness should tap into whether the drill practice fulfills some of the student's intrinsic and extrinsic needs, irrespective of their overall motivation for the task of learning Chinese. Indeed, analysis of subjective task value theory proposes that intrinsic value is a component of judging task value, and in this analysis, intrinsic value is closely aligned with flow theory[Eccles 2005], since flow is defined as optimal experience. If the student feels like they are not meeting some extrinsic or intrinsic goal while practicing, they will be unlikely to report the practice as useful, regardless of whether they have an overall motivation for Chinese. While one would not wish to assume flow [Csikszentmihalyi 1990] is synonymous with

usefulness, since usefulness may be a broader construct than flow including more extrinsic aspects of value, it does seem plausible to suppose that these momentary usefulness judgments measure an important aspect of flow.

The third question, about whether the student used a strategy based or a rehearsal-based procedure during the recent practice is intended to be a way to differentiate the association procedure that the student used, by framing strategy use in opposition to less strategic repetition. This strategy use construct is important because of the strong effects strategies are shown to have in flashcard learning tasks such as this [Atkinson and Raugh 1975; Pavlik Jr. 2007]. This strategy use question implicitly captures the propensity of students to use prior knowledge to form associations, in contrast to forming associations with a minimum of prior knowledge involvement (repetition). Previous research suggests there are many interactions between learning and strategic knowledge for a variety of more complex tasks [e.g. Alexander and Judy 1988], and one might expect this pattern to be true for flashcard learning as well. In the context of flow theory, the strategy question asks the student to consider whether they have felt they are applying any coordinated strategy (skills) for the task. Here it might be expected that such strategy use (types of which they are exposed to, but not taught during the pre-survey) will balance challenges according to the flow theory. It is careful to highlight, however, that unlike Csikszentmihalyi's construct of prior skill in the flow domain, our strategy use question asks for learning strategies actually used during the recent practice.

Finally, there were multiple measures of drill performance during the flashcard learning. To make the project tractable at this stage, only percent correct, latency, the flashcard delivery system parameter (memory decay) that controlled the spacing of practice during learning, and the count of epochs (2 minute intervals of practice) of practice completed were analyzed. While percent correct is to some extant controlled by the optimal scheduling of flashcards, prior results have found there is a great deal of variability of correctness that cannot be easily explained by the scheduling model [Pavlik Jr. et al. 2008] (indeed, this was a main motivator to begin determining the effect of the variables in this paper that are currently not used by the practice scheduling model for scheduling decisions). Correctness during flashcard practice has been shown to improve practice efficiency [Pavlik Jr. and Anderson 2008], and one primary goal of creating a dynamical systems model was to determine whether and to what extent the three motivational and metacognitive construct variables measured appear to modulate correctness. Similarly, latency of responding is often inversely correlated with

performance in memory tasks [Judd and Glaser 1969] and may provide additional understanding of the relationships particularly because strategy use might be increased by higher latencies despite the fact that strategy use tends to improve performance [Pavlik Jr. 2007]. The decay parameter dynamics were also investigated, since higher decay in the model implies the student was being given more repetitive practice because the spacing between repetitions was reduced if the model assumed people were forgetting quickly. Finally, the effects on the duration of practice (or count of epochs) were also analyzed, which would be useful to see what conditions led to more practice time.

A dynamical system model of these constructs is useful both to understand the practice paradigm, and for the long-term potential of using such a model to react adaptively to student states. To begin understanding this potential, after describing the data preparation and construct measurement model from a factor analysis of survey results, hypotheses are tested and simulations are presented with the goal of better understanding the system found. By simulating students with various fixed or momentary tendencies, it may be possible to see how and why the tutoring system seems to fail or succeed for these sorts of students. The goal behind the modeling was to arrive at predictions of the best start states through simulation. These predictions can then be used to design experimental interventions that attempt to start students at these felicitous beginnings.

3. DATA PREPARATION

The data was collected in the Elementary Chinese II course at Carnegie Mellon University (CMU). All data used to create the model have been archived in the Pittsburgh Science of Learning Center (PSLC) DataShop web application (dataset id# 390), and a copy of the model and data are available to registered users (free). The R code posted at this location functions by downloading the data using DataShop web services (requires a quick set-up) before finding the model using R functions as described below, performing cross-validation, running simulations, and testing models of flow.

The data itself comes from Lessons 15-18 in Elementary Chinese II at Carnegie Mellon University, which correspond to chapters 15-18 in the *Chinese Link* (*Zhongwen Tiandi*) textbook [Wu et al. 2006]. The intervention began with a pre-survey and concluded with a post-survey, both delivered from surveymonkey.com. The intervention itself was a computerized adaptive practice drill system with several different types of

Chinese vocabulary flashcards including practice of the pinyin phonetic system, English meaning, radical character components of Chinese, and flashcards where the student filled in a missing vocabulary word in a sentence (Cloze fill in the blank) [Pavlik Jr. et al. 2008]. There were 55 unique students with complete data, consisting of 168 unique lesson runs across the 4 units. Average lesson-run length was about 20 minutes (more precisely there was an average of 9.39 epochs of data per lesson run, which implies 8.39 analyzable transitions since the first epoch had no prior data to allow prediction). Students were allowed to take breaks (by hitting a pause button) at any time, the dynamical systems model ignores any breaks students took and treats each lesson as continuous. Since students were asked to practice for only 20 minutes, many students completed the work without any break. While there was a between-subject manipulation of initial practice difficulty (either wider or narrower distribution of initial practice), this did not appear to cause detectable effects on learning at the pre/post-test level, though it may have produced more varying data, thus providing more variability for the model to capture, improving the detection of patterns by encouraging the occurrence of these patterns in the data. This lack of overall effect of the manipulation was not unexpected because of the relatively minimal amount of additional practice students got in the system (only 20 minutes for a 1 week long lesson), and due to the possibility that the high achieving CMU students in the Chinese course would compensate by studying outside the system if they found it less useful.

There were seven data vectors per student per lesson, including vectors for easiness, usefulness, strategy use, and performance. Since each of the 3 survey items was measured sparsely, about 2/3 of the data values in each survey vector were empty and so were filled in with 0 values. These 0 values were then manually excluded from prediction as dependents, and therefore did not influence solving for model coefficients as independent variables. Probability correct for the previous 2 minutes (previous epoch) however, was always calculable, so these vectors were not sparse. Further, since probability (performance) is a 0 to 1 value (tending to be greater than 50% in the data), whereas the readings from the other measures varied from 1-5, probability was scaled by using the logit of probability as data (constrained so that logit $< -3 \rightarrow -3$ and logit $> 5 \rightarrow 5$). For this reason, the model predicts logit values for performance, and the simulator code converts these to probabilities for the simulation graph plotting in the discussion. Latency and decay were computed from the data in a way analogous to correctness, where latency was measured from the trial's start until the first key of the response was typed, in

seconds. Decay was measured from 0.1 to 0.3 with 0.3 indicating fast decay, which results in narrower spacing in the system to compensate for the presumably faster forgetting. Count of prior epochs was computed starting with 0 for the first epoch for each lesson for each user.

4. MEASUREMENT MODEL

An exploratory factor analysis measurement model on the measures on the pre-survey was computed. The goal of this factor analysis was to provide a number of covariates that could be used to better see and understand the dynamic effects in the data. Furthermore, because of the considerable overlap between the pre-survey factors (to be describe below) and the Likert items during practice, the interpretation of the dynamic factors as dynamic states, to the extent they are still supported, is enhanced since the effect of the more constant trait like pre-survey factors is controlled for. Table 1 lists the subjective construct label for each of the factors that were found.

The items input to the factor analysis are shown in Table 1, split into the factors the analysis found. There were 2 parts to the pre-survey. The first part of the pre-survey had 6 strategy use items created by the experimenter. Items on the strategy survey were on a 5-point Likert like scale with 1=I have never heard of this strategy and 5=I have used this strategy whenever possible. The final 16 items on the survey were taken from a 4 factor abbreviated version of the MSLQ [Pintrich et al. 1993] that was selected based on prior usage of the full scale in similar populations [Pavlik Jr. 2010]. This abbreviated version used the most reliable items of the strongest factors from this prior work. Items on the MSLQ were standard 7-point Likert scale items. The factor analysis found 7 factors with eigenvalues greater than 1. These 7 factors explained 72.7% of variance and were extracted with principal components analysis and subject to a varimax rotation that converged after 8 rotations. Factor scores were computed with the SPSS regression method for entry into the subsequent dynamical systems by student.

Table 1. Factors found in the survey, with respective factor loadings for each item in factor.

Factor 1 – Subjective label: positive value
Understanding the subject matter of this course is very important to me. 0.802
I am very interested in the content area of this course. 0.812
It is important for me to learn the course material in this class. 0.768
I like the subject matter of this course. 0.813

Factor 2 – Subjective label: efficacy for class I'm confident I can understand the most complex material presented by the instructor in this course. 0.909 I believe I will receive an excellent grade in this class. 0.530 I'm confident I can do an excellent job on the assignments and tests in this course. 0.683 I'm certain I can understand the most difficult material presented in the readings for this course. 0.530
Factor 3 – Subjective label: anxiety I feel my heart beating fast when I take an exam. 0.565 I have an uneasy, upset feeling when I take an exam. 0.697 When I take a test I think about how poorly I am doing compared with other students. 0.841 When I take tests I think of the consequences of failing. 0.782
Factor 4 – Subjective label: self-regulated learning When I study the readings for this course, I outline the material to help me organize my thoughts. 0.679 When I study for this class, I pull together information from different sources, such as lectures, readings, and discussions. 0.687 I make simple charts, diagrams, or tables to help me organize course material. 0.829 If I get confused taking notes in class, I make sure I sort it out afterwards. 0.522
Factor 5 – Subjective label: verbal mediation strategy Sound symbolism - in this strategy you link the Chinese word by considering how it sounds and linking that with the meaning 0.783 Verbal Link - in this strategy you link the Chinese word by trying to think of it in a sentence or by using it in a short phrase 0.804 Radical phonetic meaning – in this strategy you use the pinyin sound of the radical as a clue to the word meaning 0.610
Factor 6 – Subjective label: substitution strategies Keyword link - in this strategy you think of a native language word that sounds like the Chinese word and then you link this similar sounding word with the meaning 0.8341 Radical phonetic pronunciation – in this strategy you use the pinyin sound of the radical as a clue to the word pronunciation 0.740
Factor 7 – Subjective label: imagery strategy Image symbolism - in this strategy for learning Hanzi, you look at the character and try to link the meaning with some part of the Hanzi character's form, such as a radical 0.739

5. DYNAMICAL SYSTEM MODEL

A dynamical systems model assumes some current state of nature and then describes an evolution rule about how each state is transformed to the future state [Ward 2002]. Assuming the evolution rule is correct; by iteration, we can predict the future states of nature given some start state. Dynamical systems can be described in terms of differential equations, which capture change in terms of continuous time, but dynamical systems can also be described in discrete time using difference equations, which are equivalent to standard multiple regression methods. Because the method for collecting data only provides data for discrete 2-minute epochs and not continuously, difference equations were particularly appropriate. This was also an advantage because difference equations are mathematically simpler than differential equations.

The evolution rule was found by using linear multiple regression with and without the measurement model factors added in as covariates by subject. For each dependent, the model was simplified using stepwise regression. The backwards stepwise algorithm (from R) used here searches from the full model subtracting variables according to AIC model fit tests. This simplification step also makes the models much easier to interpret, since unnecessary non-significant model terms are automatically removed. Because the models shown are not based on strong prior hypotheses, this work is classical data mining and so suffers from the problems of over-fitting common to many model-fitting problems. 10-fold cross-validation was used to measure the extent of over-fitting that comes because of the procedures above. The cross-validation procedure protects against rejecting the null hypothesis of no effect due to over-fitting the data by training the model with 90% of the students, and then testing on the remaining 10%, in an iterative procedure that tests each 10% data fold once. Furthermore, because of the auto-correlative component in some of the models there was a risk that the cross-validations might simply be confirming the generalizability of the autoregressive component. To understand this risk, the cross-validated models without the autoregressive component are also presented for cases where an autoregressive component is included. These models provide a more rigorous test of the fit under conditions where the autoregressive component (though perhaps valid) is unable to inflate the model correlations. Cross-validation was computed for both the full models with all linear terms, and the stepwise-reduced linear model. Models with all interactions, and 2nd order squared components as reported previously [Pavlik Jr. and Wu 2011], were found to fit poorly after stepwise reduction when cross-validation was computed and are not considered here (in addition to

the fact that the data preparation in these prior results biased the model). All cross-validations are the average correlations for 10 runs of a 10-fold cross-validation, with data selected into folds by student.

6. RESULTS

The evolution rules will be shown for the dependents easiness, usefulness, strategy use, performance, latency, the optimization system decay parameter and the count of epochs for the session. The models predicting each of these includes the prior epoch's dependents, serving as independents, and the 7 factors for the subject from the measurement model. Tables 2-8 show the stepwise-reduced parameter set with coefficients and p-values from the multiple regressions. Model 1 and 2 columns include the measurement model values for each subject in the prediction of the dynamical system for each subject. Model 1 and 3 columns include an autoregressive component, which means that the prior epoch's value for that variable was also included in the calculation. Autoregressive models were not computed for the easiness, usefulness, and strategy use variables since they were only gathered every third epoch. Additionally, the autoregressive models were not computed for the count variable prediction because count was perfectly predictive because it simply increases by one each epoch. When interpreting the parameters and models, it is useful to see which parameters are consistent across the 2 or 4 models for each dependent, since these are clearly the strongest effects. Further, during interpretation, it is also useful to examine the cross-validation correlations, since depending on the dependent measure; some of the models are relatively poorly fit.

In both models of the dependent easiness, there are significant effects showing that students that report high strategy use, usefulness and performance subsequently report experiencing more easiness. See Table 2. Interestingly, there appears to be some negative relation involving several of the measurement model factors and easiness in Model 2. However, the results for Model 2 are somewhat questionable given the poor cross-validation of this model, which implies model fit may have been biased by outlier data and is therefore not useful. Further, Model 4's relatively poor fit diminishes this result, which nevertheless is uncontroversial since it indicates that correctness, low latency and narrow spacing (higher decay) are all associated with easiness. Finally, strategy use and usefulness are relatively strong predictors of easiness ratings.

Table 2. Easiness predictor list showing the linear coefficients with *p-values*.

	Model 2	Model 4
intercept	2.2132, p=6.11e-30	2.3371, p=9.22e-30
easy		
useful	0.1861, p=2.20e-05	0.1606, p=1.77e-04
strategy	0.2080, p=3.40e-06	0.2040, p=4.88e-06
performance	0.0584, p=4.95e-02	0.0602, p=4.46e-02
latency		-0.0427, p=1.35e-01
decay	1.6236, p=2.78e-02	1.8421, p=1.25e-02
count		
value		
efficacy		
anxiety	-0.1064, p=1.65e-02	
self-regualted	-0.0987, p=3.02e-02	
verbal mediation	-0.0713, p=9.72e-02	
substitution		
imagery	-0.0837, p=6.23e-02	
r for stepwise model	0.2946	0.257
r - 10 x 10-fold CV stepwise	-0.0359	0.155
r - 10 x 10-fold CV full	-0.0313	0.181

Usefulness of the procedure as reported by users during practice was relatively well predicted, as shown in Table 3. In both models, there were effects of strategy use and easiness predicting subsequent reports of usefulness. Further, it appears that students who are slower to respond report higher usefulness, while students with lower decay (wider spacing) also find the practice more useful. While the measurement model components of Model 2 reach significance, the poorer fit compared to Model 4 on cross validation implies that these results are not trustworthy, particularly since Model 4 has a better correlation despite half as many parameters.

Table 3. Usefulness of learning predictor list showing the linear coefficients with *p-values*.

	Model 2	Model 4
intercept	1.8635, p=4.93e-20	1.625, p=8.40e-16
easy	0.3844, p=5.71e-18	0.430, p=6.48e-21
useful		
strategy	0.4157, p=9.07e-17	0.469, p=6.07e-20
performance		
latency	0.0796, p=2.13e-02	0.112, p=1.69e-03
decay	-2.1517, p=3.63e-03	-2.391, p=1.91e-03
count		
value	0.1834, p=1.16e-05	
efficacy	-0.1618, p=3.32e-04	
anxiety	0.0986, p=2.04e-02	
self-regualted	0.1193, p=8.55e-03	
verbal mediation		
substitution		
imagery		
r for stepwise model	0.539	0.464
r - 10 x 10-fold CV stepwise	0.271	0.323
r - 10 x 10-fold CV full	0.271	0.315

Strategy reports were best predicted by prior usefulness and easiness. See Table 4. One might be concerned whether the items for these three self-reports, strategy use, easiness and usefulness of learning, are really tapping the same construct, e.g. propensity to engage; however, on the surface the strategy and usefulness items ask quite different questions. Also useful to help answer this question is to note that while the prediction pattern for both decay and latency was the same for usefulness and strategy use, it was reversed for easiness, implying a qualitative difference in what easiness ratings measured relative to usefulness and strategy use ratings. Other predictors also differ, for instance, higher performance seems to predict easiness and strategy use, but no effect was found for usefulness. Again, it appears with the extra factors from the measurement model we have an over-fit situation since cross-validation indicates the Model 2 generalizes less well that Model 4.

Table 4. Strategy use predictor list showng the linear coefficients with *p-values*.

	Model 2	Model 4
intercept	2.0486, p=2.24e-29	1.8266, p=4.42e-20
easy	0.2717, p=2.11e-11	0.2733, p=5.91e-11
useful	0.3231, p=1.67e-14	0.3270, p=2.68e-14
strategy		
performance		0.0505, p=9.53e-02
latency		0.0730, p=1.37e-02
decay	-1.2296, p=7.76e-02	-1.8970, p=7.39e-03
count		
value	0.1325, p=7.92e-04	
efficacy	-0.1199, p=3.52e-03	
anxiety	0.1189, p=3.67e-03	
self-regualted		
verbal mediation		
substitution		
imagery	0.0957, p=2.75e-02	
r for stepwise model	0.437	0.377
r - 10 x 10-fold CV stepwise	0.177	0.25
r - 10 x 10-fold CV full	0.158	0.253

Performance prediction in Models 2 and 4 was in line with the previous variables, but since we gathered this measure continuously, we were able to produce Models 1 and 3 with autoregressive components, see Table 5. Except for Model 2, there appears to be an effect of strategy reports being somewhat predictive of later performance. Similarly, in the pre-survey, we find that higher values on the substitution factor also predicted better performance. Other pre-survey effects included a tendency for higher performance as a function of the value factor and lower performance as a function of the efficacy factor. Higher efficacy may have this negative effect if it leads to carelessness during the practice (an overconfidence effect). As we might expect, narrower spacing (caused by higher decay) and more practice (count) correlated with higher performance.

Table 5. Performance predictor list showing the linear coefficients with *p-values*.

	Model 1	Model 2	Model 3	Model 4
intercept	0.5478, p=6.79e-06	0.9486, p=1.48e-12	0.4982, p=4.60e-05	0.6424, p=3.62e-04
easy				0.0977, p=1.80e-02
useful				0.0824, p=5.33e-02
strategy	0.0376, p=1.02e-01		0.0436, p=6.10e-02	0.1266, p=5.45e-03
performance	0.4909, p=1.33e-76		0.5268, p=1.78e-89	
latency				
decay	1.2519, p=3.14e-02	2.8075, p=1.88e-05	1.1571, p=4.82e-02	2.7596, p=3.95e-05
count	0.0516, p=1.88e-06	0.1071, p=6.75e-19	0.0464, p=1.95e-05	0.1015, p=1.35e-16
value	0.1206, p=2.36e-04	0.2145, p=5.81e-09		
efficacy	-0.1037, p=3.03e-03	-0.1950, p=7.17e-07		
anxiety		-0.0572, p=1.29e-01		
self-regualted		0.0591, p=1.52e-01		
verbal mediation				
substitution	0.1065, p=1.29e-03	0.1798, p=1.96e-06		
imagery				
r for stepwise model	0.553	0.339	0.537	0.237
r - 10 x 10-fold CV stepwise	0.494	0.201	0.498	0.202
r - 10 x 10-fold CV full	0.492	0.201	0.499	0.206

Latency was predicted with somewhat less accuracy than performance, and we can see that without the autoregressive component, cross-validation was very poor indicating catastrophic over-fit. See Table 6. This seems unsurprising, since individual differences in latency of responding would be expected for different strategies, prior knowledge, or ability in the task. Of particular note was the very strong effect of anxiety in Models 1 and 2 on latency, which shows a clear effect of longer latencies when anxiety was higher, an interesting effect that has been found previously for semantic memory[Hartley et al. 1982]. More expected was the significant effect of more practice (count) on reducing latency. Additionally, Model 3 showed several marginal effects (p>.05) that nonetheless were in the expected directions.

Table 6. Latency predictor list showing the linear coefficients with *p-values*.

	Model 1	Model 2	Model 3	Model 4
intercept	1.9552, p=8.50e-88	3.3180, p=0.00e+00	1.7507, p=1.72e-31	3.3299, p=0.0000
easy			0.0478, p=1.81e-01	
useful			0.0382, p=3.01e-01	
strategy			0.0528, p=1.81e-01	
performance	-0.0361, p=1.17e-01		-0.0290, p=2.21e-01	
latency	0.4027, p=1.82e-61		0.4303, p=1.12e-69	
decay			-0.1093, p=7.87e-01	
count		-0.0227, p=4.21e-02	-0.0065, p=5.72e-01	-0.0255, p=0.0255
value	0.0472, p=1.41e-01	0.0693, p=4.69e-02		
efficacy	-0.0624, p=6.98e-02	-0.0829, p=2.72e-02		
anxiety	0.1815, p=7.14e-08	0.2971, p=4.08e-16		
self-regualted				
verbal mediation				
substitution				
imagery				
r for stepwise model	0.466	0.2347	0.447	0.05874
r - 10 x 10-fold CV stepwise	0.384	0.092	0.394	0.00601
r - 10 x 10-fold CV full	0.372	0.0707	0.391	-0.00184

The most stable predictor of decay was prior decay (autoregression) as shown in Table 7. There was some suggestion that using imagery might have a negative effect on decay, perhaps if imagery strategies result in high performance, this might occur because of the adaption of the system with wider spacing of practice. However, since this did not occur for performance, one might suspect Model 2 is over-fit as suggested by the cross-validation, which shows that it fits worse than the other 3 models. Without the autoregressive component, there was good performance leading to narrower spacing, which is anomalous, but may be explained by the relative insensitivity of the decay parameter, which only moved in small steps to automatically calibrate. Because of this, the positive correlation between narrow spacing (high decay) and performance was picked up in this analysis, rather than the longer run tendency to for the system to have subsequently lower decay if performance is good. It may be useful to note that decay predictions may be less useful than other measures because of the tendency for decay to decrease across practice because of the initial calibration. This meant that decay correlated with epoch count and may have been predicted by variables that correlated with the count.

Table 7. Decay predictor list showing the linear coefficients with *p-values*.

	Model 1	Model 2	Model 3	Model 4
intercept	0.00320, p=2.46e-01	0.14175, p=7.16e-100	0.00202, p=4.62e-01	0.14172, p=9.08e-100
easy	-0.00130, p=2.52e-02	-0.00226, p= 1.39e-01	-0.00133, p=2.28e-02	-0.00320, p= 4.02e-02
useful	-0.00143, p=1.74e-02	-0.00278, p= 8.01e-02	-0.00148, p=1.45e-02	-0.00405, p= 1.20e-02
strategy	-0.00141, p=2.79e-02	-0.00374, p= 2.71e-02	-0.00146, p=2.33e-02	-0.00504, p= 3.44e-03
performance		0.00530, p= 4.53e-07		0.00510, p= 1.51e-06
latency	0.00224, p=1.73e-08	0.00836, p= 1.25e-15	0.00225, p=1.78e-08	0.00909, p= 9.10e-18
decay	0.89396, p=0.00e+00		0.90037, p=0.00e+00	
count	0.00130, p=6.50e-14	-0.00238, p= 1.46e-07	0.00130, p=8.92e-14	-0.00231, p= 6.10e-07
value		-0.00366, p= 9.58e-03		
efficacy				
anxiety		0.00326, p= 2.49e-02		
self-regualted		-0.00357, p= 2.25e-02		
verbal mediation		0.00328, p= 2.23e-02		
substitution		0.00418, p= 3.73e-03		
imagery	-0.00181, p=1.54e-03	-0.01261, p= 1.89e-17		
r for stepwise model	0.934	0.384	0.934	0.29
r - 10 x 10-fold CV stepwise	0.923	0.198	0.923	0.248
r - 10 x 10-fold CV full	0.922	0.21	0.923	0.251

The count of trials dependent variable could also be computed from the independent components, however, in this case, Model 1 and Model 3 that included the autoregressive component were perfectly fit, since prior time increments correlated perfectly with future time increments. Table 8 shows only Models 2 and 4. Performance led to higher counts, and narrower spacing (high decay) led to lower counts. Further, considering measurement model factors, the results showed that self-regulation had a relatively strong negative effect on counts while prior efficacy led to higher counts. Note that usefulness did not reflect on the count of practices. Rather count was paradoxically predicted by both better performance, faster responding and wider spacing (lower decay). As noted for decay, this may have been driven by a general tendency to increase spacing as practice progressed. This tendency means that the flashcard system may have started out with too narrow spacing on average, and this resulted in continuous small adjustments as practice progressed. Nevertheless, it is difficult to dismiss the strong effect of the self-regulation factor on reducing the count. Also significant was the higher count predicted for students with high efficacy.

Table 8. Count predictor list showing the linear coefficients with *p-values*.

	Model 2	Model 4
intercept	6.568, p=2.39e-90	6.712, p=1.39e-93
easy		
useful		
strategy		
performance	0.564, p=8.32e-21	0.522, p=1.29e-18
latency	-0.131, p=2.72e-02	-0.153, p=1.07e-02
decay	-12.323, p=7.08e-18	-12.044, p=7.23e-17
count		
value	-0.148, p=6.62e-02	
efficacy	0.265, p=1.97e-03	
anxiety		
self-regualted	-0.470, p=1.13e-07	
verbal mediation		
substitution	0.128, p=1.19e-01	
imagery		
r for stepwise model	0.363	0.322
r - 10 x 10-fold CV stepwise	0.3	0.321
r - 10 x 10-fold CV full	0.3	0.314

7. DISCUSSION

One way to characterize the Csikszentmihalyi's theory of flow [Csikszentmihalyi 1990] in terms of the work here is to say that the flow theory predicts usefulness during learning should show an inverted-U shaped relationship with easiness, with high easiness

and low easiness predicting less usefulness. Indeed, *this specific model* is confirmed in the data when only examining the effect of easiness and easiness^2 (either with or without the autocorrelation from prior usefulness) on usefulness (but with an adjusted r^2 of .217). This model predicts a significant positive linear effects (p= 1.47e-13) and significant negative squared term effects (p=7.97e-12), which indicates an inverted U-shape as Csikszentmihalyi predicts. However, data was gathered on several constructs that are relevant to flow theory, such as strategy use, which parallels the importance of skill in the task as a balance for the difficulty of the task, according to flow theory. We might therefore expect a similar inverted-U shaped effect of strategy use on usefulness, since flow theory implies that as skill reaches some balanced amount, the task is found to be optimal in terms of usefulness or optimality of the learning experience. However, testing this testing this model for the effect of strategy and strategy^2 on usefulness shows no indication of an inverted-U, in contrast, there is a non-significant tendency for the squared term to be positive, indicating that, at least for the task in this paper, high levels of strategy use did not appear to have negative consequences for the experienced usefulness of the learning.

However, flow theory tends to focus on the combination of skill and challenge, rather than either variable taken alone. Therefore, elaborating on the prior analyses, it is also possible to test the flow-inspired model where usefulness is predicted by the difference between challenge (inversion of the easiness measure) and strategy use. (To do this the data was recoded so that each rating of easiness, usefulness and strategy use persisted for an additional epoch.) While this model had a fit of only adjusted r^2 equal to 0.031, this model results in a negative effect of this difference, which is significant, p=.00044. Figure 1 plots this relationship, which shows the flow theory prediction that there is a ridgeline of optimal usefulness where the difference between challenge and skill is minimized [Csikszentmihalyi 1990, p. 74]. While this result is interesting, the close match to flow theory of this model is minimized if one also includes challenge and strategy use in the model (instead of only their absolute difference). With this more complex model, the results indicate that strategy use dominates the prediction, and there is no longer a significant effect for challenge, or the difference of challenge and strategy use. This model seems to break the flow theory, since there is no optimality for a balance of strategy and challenge.

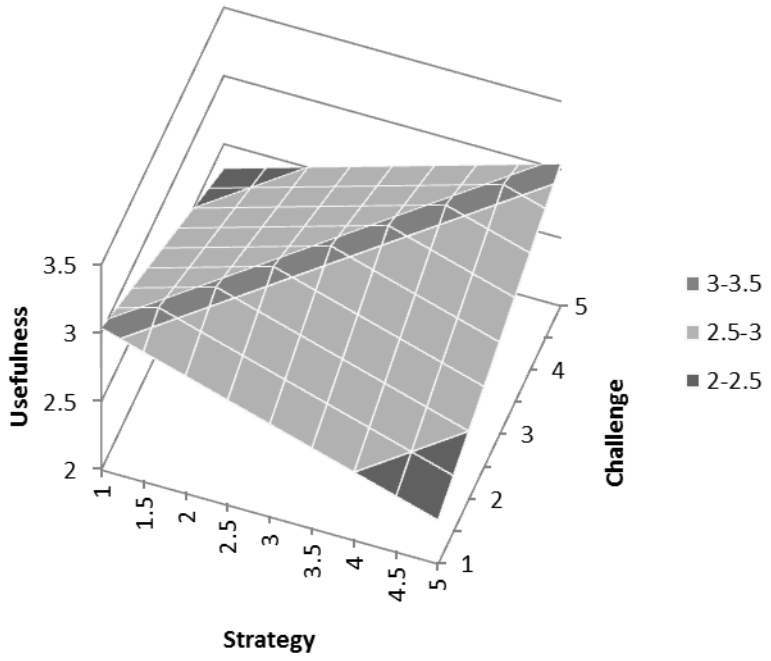


Figure 1. Usefulness surface predicted by the absolute difference between challenge and strategy use.

Digging still deeper we can refine the flow theory by specifying that it might be the squared difference rather than the absolute difference that results in low usefulness (e.g. high challenge with low skill). This refinement (a model with challenge, strategy use, and their squared difference as predictors of usefulness) implies that only extreme differences in challenge and skills will have serious effects on usefulness. This model fits relatively well, with an overall adjusted r^2 for the model of 0.34. See Figure 2, which illustrates this model. Inspecting the model shows that it provides at least tentative support for the flow theory, since the squared difference of challenge and skill has a significantly negative coefficient. However, even with this squared component added to the model, a linear effect of strategy use still dominates the prediction of usefulness for the task, since the p-value for the difference squared term only barely reached significance p=.046. In addition to the strong dominance of strategy use in predicting usefulness, one implication of this model is that when flow theory is applied to optimal learning experience there may be *an upward slope* to the flow channel such that a more optimal learning experience is had to

the extent there is a reasonable balance of both more challenge and more skill use. Interestingly, this upward slope does not appear to be a strong component of early flow theory [Csikszentmihalyi 1990, p. 74], but instead appears to be an additional complexity to the theory that has been added as it was developed [Csikszentmihalyi 1997], since this later work explicitly includes the upward slope of the flow figure by indicating that both high challenge simultaneous with high skill are necessary for maximal flow.

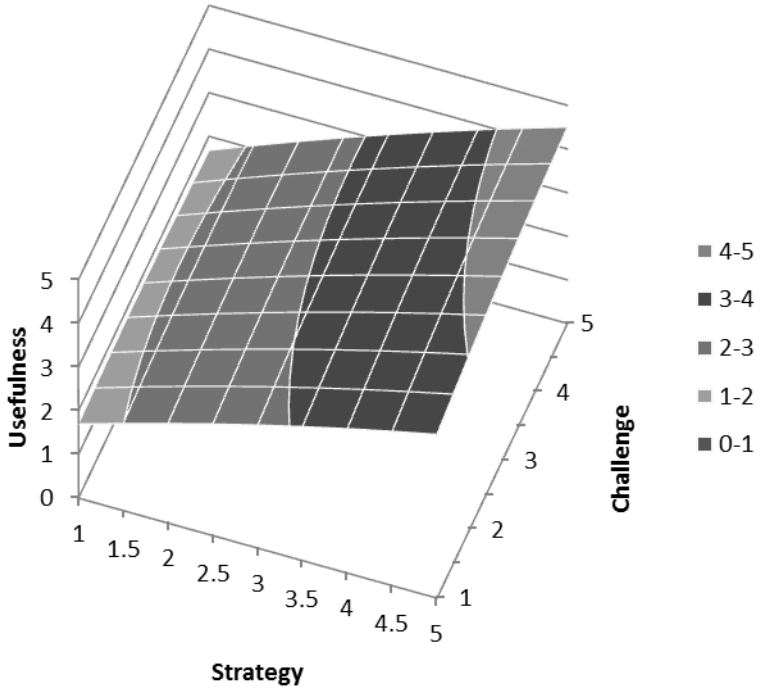
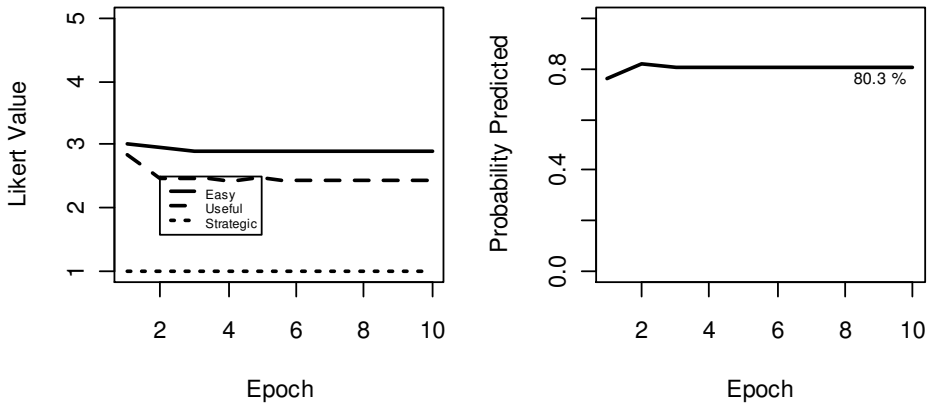


Figure 2. Usefulness surface predicted by the challenge, strategy use, the absolute difference between challenge and strategy use squared.

The dynamical systems model also appears to have potential as a way of tracing student motivation and metacognition during learning in an adaptive system, but at this stage of the research we wanted to begin by using the model as a simulation tool to show us how to understand the student experiences in the tutor and how these experiences are reflected in the constructs we tracked. These simulations can guide hypothesis testing, for instance, we can test the question of what the effect is of a usefulness manipulation as opposed to a strategy use manipulation. Below we show simulation where we "peg" or fix one or more initial values for the model and then let it run from this point to simulate

a student that begins with some initial propensity toward the system (e.g. use strategies). In these simulations, the parameters from Model 4 were used in all cases, and the first epoch was initialized at the student averages for the first epoch from the data.

Figure 3 shows the procedure of "pegging" a particular dynamic variable to some fixed value for Model 4. For instance, imagine a simulated student pegged to high strategy use has a habit of using vocabulary-learning strategies (probably developed with long experience), while a simulated student pegged to low strategy use assumes vocabulary learning is a repetitive procedure and never attempts strategies. Figure 3, top shows the case of low strategy use simulate student, and Figure 3, bottom shows the case of high strategy use. The simulation shows how students who consistently report low strategy use have poor performance, find the task harder, and find the task less useful. In particular, high strategy use has an effect on usefulness while resulting in only slight easiness effects. This might be interpreted to suggest high strategy use ratings, particularly in combination with a high usefulness rating, predict better performance, but entail more student effort and attention to the task, thus preventing it from becoming too easy. The simulated effect shows that the percentage of errors decreases from 19.7% (1-.80.3) to 15.5% (1-.84.5), a difference that may be important because it reveals almost 21% fewer errors in the low strategy case. Because errors greatly reduce practice efficiency in flashcard type learning [Pavlik Jr. and Anderson 2008], such an effect suggests strategy use may be an important component of optimal practice.



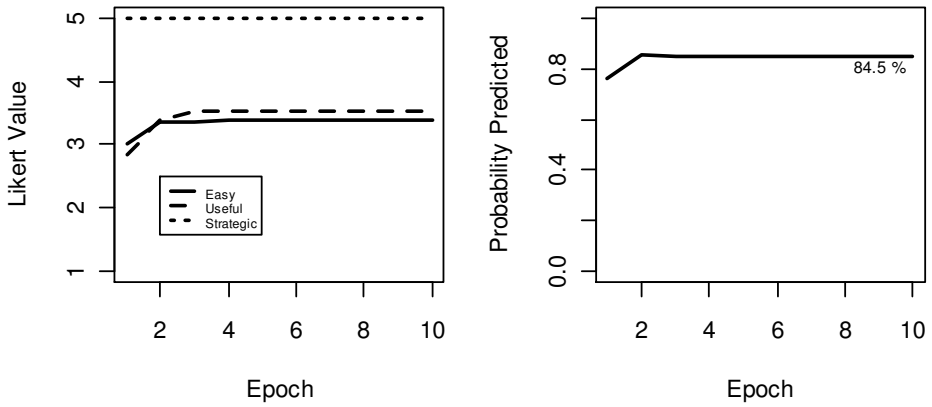


Figure 3. Pegged simulation of student who is a committed rote learner vs. strategic learner. (Plotting only 4 of the 7 dependents.)

The prior example was illustrative, but we might not necessarily expect any manipulation would easily peg real students to such a high strategy use. Instead, a similar but more realistic situation can also be created in which a transient manipulation in strategy use is simulated. Figure 4, top shows a model where a simulated student is pinned to not use strategies during the first 2 minutes of practice. The figure shows how the simulated student rather quickly reverts to the baselines in such a case. This result suggests that despite the difficulty of pegging students to new baseline levels of strategy use, such a change may be necessary to produce a long-lasting effect. Perhaps we could achieve this change through some built in assistance that makes the task easier, which has been referred to as scaffolding [Hmelo-Silver et al. 2007]. However, the model suggests that unless that strategy use is explicitly pegged (which would be caused by long-term retention of strategy behaviors combined with the motivation to use them) we cannot expect that there will be effects beyond a temporary boost in learning during the scaffolded activity.

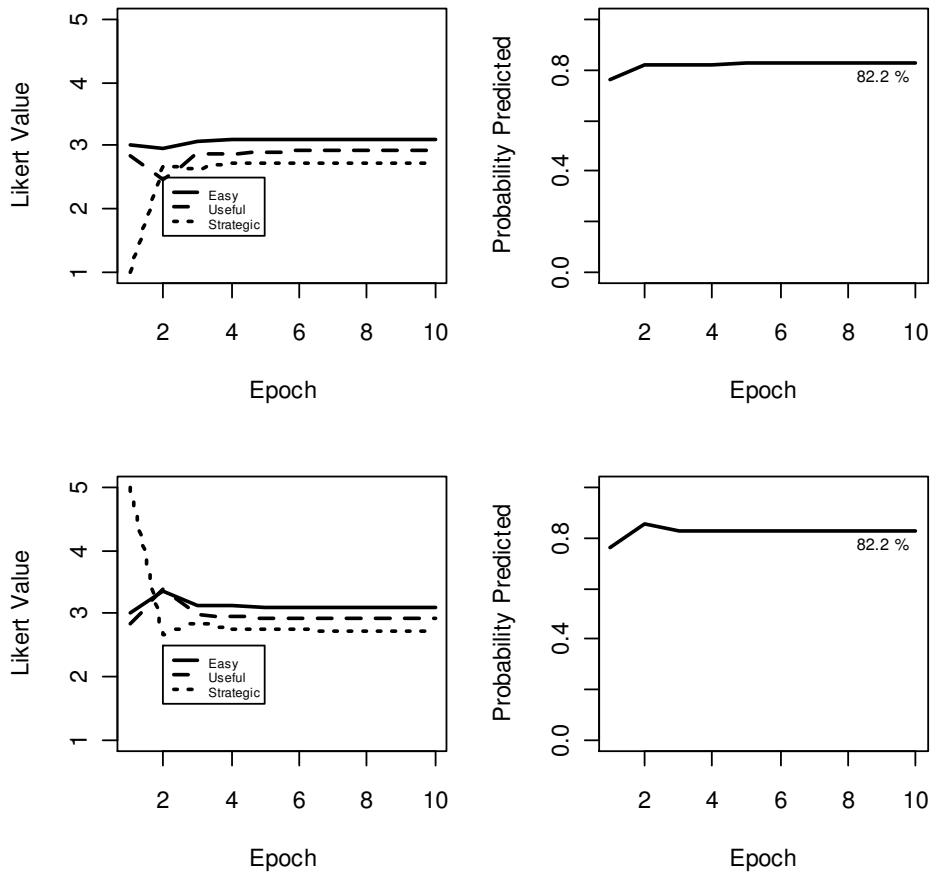


Figure 4. Pinned simulation of student who begins as rote learner vs. strategic learner. (Plotting only 4 of the 7 dependents.)

This work begs the question of how difficult it would be to peg, or cause permanent change in, this motivational and metacognitive construct system in real students. One might hypothesize, for example, that if students were given prior training in strategy use, the parameters of the strategy component of their individual dynamical system (e.g. perhaps modeled by an increase in the intercept parameter for each student's strategy evolution rule) may be modified [Rodrâiguez and Sadoski 2000; Wang and Thomas 1995; Wieczynski and Blick 1996]. This sort of change would result in students going down the virtuous bifurcation leading to high performance and feeling one's work is useful. Similar results have been shown for an educational software for computer literacy, in which measurements of flow tended to be more likely to be followed by engagement

[D'Mello et al. 2007]. Indeed, the relationships discussed above imply that we might benefit from teaching our students' vocabulary learning strategies, but they also highlight the importance of exposing the learners more directly to the potential usefulness of the learning exercises. For instance, the literature shows that repeated testing causes long-term learning [Karpicke and Roediger 2008; Thompson et al. 1978], and we might suppose that if this data was presented convincingly to students, they may find the work more useful with such justification. Such an experimental manipulation might be expected to improve usefulness ratings, and might be further focused on disambiguating the high correlation we saw between usefulness and strategy use. In other words, is it true that strategy use causes an increase in usefulness, or might this association imply that students who find the work useful are motivated to engage in strategy use more frequently?

However, the possibility that knowledge of useful strategies is sufficient is also minimized if we look at how poorly the prior knowledge of strategies from the survey measurement model predicted usefulness in contrast to prior findings of actual usefulness for mnemonic strategies when applied in a controlled fashion [Atkinson 1975; Bower 1970; Wang and Thomas 1995]. It may not be enough that the students know strategies if they are not aware of how useful these strategies are to specific contexts of learning. Indeed, it appears that strategy knowledge may need to be enacted for it to provide the beneficial boost in perceived usefulness we might desire to see. Perhaps by providing students with examples of how to use learning strategies in the context of the flashcard learning system they might be helped to see the usefulness of prior strategies (Factors 5 and 6), which currently did not predict perceived usefulness. This might be done, for example by providing strategic affordances such as premade linking contexts or keywords [Atkinson and Raugh 1975], in addition to explicit prior instruction in strategic techniques for learning. Therefore, while some students are able to report high strategy use and reap the benefits of this strategy use according to our data, we also see that sometimes strategic knowledge is inert (as suggested by the lack of effect of prior strategy factors from the measurement model factors computed from the pre-survey). It seems likely that to address these problems a combined approach that both boosts strategy knowledge and scaffolds and encourages strategy use may be necessary. Some students may need the encouragement to use the strategies they know, other students may need to build the strategies, and still other students may need help in both areas.

Further, our results appear to conflict with flow theory to the extent that flow theory proposes boredom is a consequence of high strategic knowledge (task skill). In contrast, we did not find that people reporting high levels of strategy use suffered in terms of their ratings of usefulness. Our results may contradict this claim of the flow theory perhaps because in our task (and likely in many educational tasks) it does not matter whether it is too hard or too easy, a maximal use of strategies is still optimal. To provide an analogy, if you give a math student 200 addition problems to do, there may be some boredom due to an excessive easiness of the task, but it will still be true that the student will likely find it most useful to use the most efficient strategies for mental arithmetic possible. It seems unlikely that they would find the task more interesting if they had poor mental arithmetic strategies. Similarly, when faced with the task of learning a collection of vocabulary words, it will not be a disadvantage if one is an expert memorizer. Rather, the expert memorizer is able to learn with less time for each practice in addition to learning more from each practice event regardless of the inherent pacing of the practice and other difficulty variables such as richness of the cues for recall.

However, behind this conflict between the results here and flow theory might be the fact that Csikszentmihalyi frames his theory in terms of the balance between skill and challenge, and often frames his examples in terms of automatized procedure [Rickard 1997] or perceptual motor skills, e.g. rock climbing, that may be categorically different than metacognitive strategy use. It may be that perceptual motor skills due to their non-cognitive aspect as automatized routines more clearly show the balance Csikszentmihalyi proposes. Expanding this idea, we might speculate that until strategies are automatized, or, until they become rote, more or better strategies to accomplish a task are always useful, particularly for learning since the strategy practice is not just instrumental to performance, but is inherently valued for its direct effect on improving later strategy use. In contrast, when a skill has become automatized, more facility with the skill does not make its application more useful, but rather at some point it becomes trivial and boring.

Indeed this need for more and better cognitive strategies in many educational tasks, particularly those activities that involve “deep learning”, may mean that a Vygotskian theoretical approach provides a better lens through which to view this result, and perhaps other forms of verbally mediated learning, because Vygotsky’s theory has been used to explain how people develop skill in thinking out loud through social interactions with others. Cognitive development in this theory is then described as the result of internalization of this self-talk [Vygotsky 1986]. Since mnemonic strategies in

vocabulary learning are often verbal [Crutcher and Ericsson 2000] it is reasonable to propose that individuals' failure to use strategies in the vocabulary task in this paper was a consequence of poor development of this self-talk, at least when such self-talk involves mediated strategies for simple recall.

In trying to apply the concepts of Vygotsky to computerized learning, Luckin has proposed two constructs that help us understand the issues. First, she describes the Zone of Available Assistance, which is the variety of possible means through which the system can address possible student problems by providing assistance. Second is the Zone of Proximal Adjustment, which is the adjustment of assistance that occurs based on the students' needs. So then, in this interpretation, the ZPA needs to be selected from the ZAA to maximize the students' presence in the ZPD, where learning is maximal [Luckin and du Boulay 1999].

This analytical framework maps well to the results in this paper. First we can note that the ZAA in the FaCT system only involves variation in the spacing of practice (narrower and wider), which directly implies that the zone of proximal adjustment can only accommodate students' needs in this limited scheduling based fashion. However, some students sometimes do not find themselves in the ZPD with such adjustment. Apparently (according to the data here), these students have not developed the mediational strategies necessary to make the task feel useful given the level of effort. Unfortunately our current system does not scaffold these strategies (include them in the ZPA) so these students find the task harder and less useful. Looking at the initial survey of strategy experience suggest this is a plausible conclusion since 16 of 60 subjects taking the pretest indicated a maximum of 3 of 5 ("I have used this strategy occasionally") across the 6 strategy items forming Factors 5, 6 and 7 in the measurement model. This indicates 25% of the students may have been only occasional strategy users even of their best strategies. Even if we assume this estimate is accurate, (and not an overestimate of strategic knowledge) it is clear a large segment of the students was poorly prepared for the intensive vocabulary practice that was provided.

In summary then, the data provide only tenuous support for the idea that optimal experience or usefulness is maximized when challenge and strategy-use are balanced. While we do see significant effects along these lines, many of these effects are small in magnitude. In contrast to the flow theory, we find that the virtuous cycle between usefulness and strategy use seems a more accurate way to summarize these results. So while we have shown effects of extreme easiness on reducing usefulness, and some effect

of balanced easiness and strategy use being optimal, all of the models show support the conclusion that easiness effects are dwarfed by the effects of tight correlation between strategy use and usefulness. To understand this effect it is useful to consider how it reflects differences in student preparedness (where they are currently at in their ZPD), with less prepared students finding the vocabulary exercises less useful since they had not developed strategies in their ZPD and the system's ZPA did not address this lack.

One limitation of this work is that it was conducted with a population of students at a high tier educational institution, and so we might suppose that this bias may make these results less general. It may be that students selected from such a situation have already become set in a motivational system where strategy use opportunities are rated as useful because of their already high value for school success. Clearly, if our students did not have this value system is would seem correspondingly less likely that strategy use and usefulness would be so tightly related. Nevertheless, given the fact that our students did produce many self-reports of poor usefulness, the high correlation we see with self-reported strategy use does provide good support for the summary conclusion above within the population studied. Regarding limitations of the work, it is also relevant to note that our measurement model factor of "self-regulated" learning (factor 4) showed a negative correlation ($r = -.32$, $p = .021$) with a follow-up item we had on a post-survey that asked "I found the frequent questions about strategies, difficulty, and usefulness during practice to be annoying." This might be taken to imply that there was the potential that the pop-up items effected performance, however, it seems such an effect might be expected to be insignificant, since there was no trend for this annoyance with the pop-up items to correlate with either overall amount of practice or performance during practice.

The main contribution of this work was to provide a general model to help us understand the dynamic motivational system of students in the task. The dynamical systems model produced is valuable because of what it implies for optimal use of the flashcard system and optimal learning in general. The model allows both testing hypotheses and simulation. As shown in the simulations, the model provides evidence of the mutually supportive feedback between usefulness, strategy use, and performance. By observing the simulations we see how high strategy use has an effect on students' perception of usefulness. These simulations allow pretesting of specific experimental manipulations and in so doing highlight how the dynamical systems model created here embodies a falsifiable theory about the domain.

8. FUTURE WORK

Future work in this area will require that researchers continue to gather such dynamic measures so that models such as the ones presented here can be further tested and confirmed. Regarding strategy use and usefulness ratings, the data indicate future interventions using similar methods might be enhanced by specific instruction in how to apply mnemonic strategies in the task. This future work would also likely be interested in gathering more measures, such as reports of engagement and enjoyment, which are supposed to correlate with flow. Furthermore, there is the opportunity to use input from momentary (or contextual) detectors of affect and motivation [Baker et al. 2008; Baker et al. 2007]. Other more objective inputs might be used, such as measures of EEG readings [Mostow et al. 2011] or measures of posture [D'mello and Graesser 2007]. Having these multiple measures probably will require that we employ some method of factor reduction for the dynamic data to better identify the constructs involved [Boots and Gordon 2011]. Finally, a continuing goal of this work is to predict longer-term transfer of learning measures.

ACKNOWLEDGEMENTS

This research was supported in part by a grant from the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation award number SBE-0836012, and Ronald Zdrojkowski for educational research.

REFERENCES

ALEXANDER, P.A. AND JUDY, J.E. 1988. The Interaction of Domain-Specific and Strategic Knowledge in Academic Performance. *Review of Educational Research 58*, 375-404.

ANDERSON, J.R., BOTHELL, D., BYRNE, M.D., DOUGLASS, S., LEBIERE, C. AND QIN, Y. 2004. An Integrated Theory of the Mind. *Psychological Review 111*, 1036-1060.

ANDERSON, J.R. AND LEBIERE, C. 1998. *The atomic components of thought*. Lawrence Erlbaum Associates, Mahwah, NJ.

ATKINSON, R.C. 1975. Mnemotechnics in second-language learning. *American Psychologist 30*, 821-828.

ATKINSON, R.C. AND RAUGH, M.R. 1975. An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning & Memory 1*, 126-133.

BAKER, R., CORBETT, A. AND ALEVEN, V. 2008. More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In *Proceedings of the 9th International Conference on Intelligent Tutoring*

Systems, B. WOOLF, E. AIMER AND R. NKAMBOU Eds. Springer-Verlag Berlin, Heidelberg, 406-415.
BAKER, R., RODRIGO, M. AND XOLOCOTZIN, U. 2007. The Dynamics of Affective Transitions in Simulation Problem-Solving Environments. In *Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction* Springer-Verlag Berlin, Heidelberg, 666-677.
BAKER, R.S.J.D., D'MELLO, S.K., RODRIGO, M.M.T. AND GRAESSER, A.C. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies 68*, 223–241.
BANDURA, A. 1997. *Self-efficacy: The exercise of control*. Worth Publishers.
BOOTS, B. AND GORDON, G.J. 2011. An Online Spectral Learning Algorithm for Partially Observable Nonlinear Dynamical Systems. In *Proceedings of the 25th Association for the Advancement of Artificial Intelligence Conference*, 293-300.
BOWER, G.H. 1970. Analysis of a mnemonic device. *American Scientist 58*, 496-510.
CRUTCHER, R.J. AND ERICSSON, K.A. 2000. The role of mediators in memory retrieval as a function of practice: Controlled mediation to direct access. *Journal of Experimental Psychology: Learning, Memory, and Cognition 26*, 1297-1317.
CSIKSZENTMIHALYI, M. 1990. *Flow: The psychology of optimal experience*. Harper and Row, New York.
CSIKSZENTMIHALYI, M. 1997. *Finding flow: The psychology of engagement with everyday life*. New York, NY, US: Basic Books.
D'MELLO, S. AND GRAESSER, A. 2007. Mind and Body: Dialogue and Posture for Affect Detection in Learning Environments. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, R. LUCKIN AND K.R. KOEDINGER Eds. IOS Press, 1563631, 161-168.
D'MELLO, S., TAYLOR, R. AND GRAESSER, A.C. 2007. Monitoring affective trajectories during complex learning. In *Proceedings of the 29th Annual Cognitive Science Society*, 203–208.
ECCLES, J.S. 2005. Subjective task value and the Eccles et al. model of achievement-related choices. In *Handbook of Competence and Motivation*, A.J. ELLIOT AND C.S. DWECK Eds. Guilford, New York, 105-121.
GOTTMAN, J., SWANSON, C. AND SWANSON, K. 2002. A general systems theory of marriage: Nonlinear difference equation modeling of marital interaction. *Personality and Social Psychology Review 6*, 326.
GUASTELLO, S.J., JOHNSON, E.A. AND RIEKE, M.L. 1999. Nonlinear Dynamics of Motivational Flow. *Nonlinear Dynamics, Psychology, and Life Sciences 3*, 259-273.
HARTLEY, L.R., SPENCER, J. AND WILLIAMSON, J. 1982. Anxiety, diazepam and retrieval from semantic memory. *Psychopharmacology 76*, 291-293.
HMELO-SILVER, C.E., DUNCAN, R.G. AND CHINN, C.A. 2007. Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist 42*, 99 - 107.
JUDD, W.A. AND GLASER, R. 1969. Response Latency as a Function of Training Method, Information Level, Acquisition, and Overlearning. *Journal of Educational Psychology 60*, 1-30.
KARPICKE, J.D. AND ROEDIGER, H.L., III 2008. The critical importance of retrieval for learning. *Science 319*, 966–968.
LIEBOVITCH, L.S., NAUDOT, V., VALLACHER, R., NOWAK, A., BUI-WRZOSINSKA, L. AND COLEMAN, P. 2008. Dynamics of two-actor cooperation-

competition conflict models. *Physica A: Statistical Mechanics and its Applications 387*, 6360-6378.
LUCKIN, R. AND DU BOULAY, B. 1999. Ecolab: The development and evaluation of a vygotskian design framework. *International Journal of Artificial Intelligence in Education 10*, 198-220.
MILLER, R.B., DEBACKER, T.K. AND GREENE, B.A. 1999. Perceived instrumentality and academics: The link to task valuing. *Journal of Instructional Psychology 26*, 250-260.
MOSTOW, J., CHANG, K.-M. AND NELSON, J. 2011. Toward Exploiting EEG Input in a Reading Tutor Artificial Intelligence in Education. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, G. BISWAS, S. BULL, J. KAY AND A. MITROVIC Eds. Springer, 230-237.
PAJARES, F. 1996. Self-Efficacy Beliefs in Academic Settings. *Review of Educational Research 66*, 543-578.
PAVLIK JR., P.I. 2007. Understanding and applying the dynamics of test practice and study practice. *Instructional Science 35*, 407–441.
PAVLIK JR., P.I. 2010. Data Reduction Methods Applied to Understanding Complex Learning Hypotheses. In *Proceedings of the 3rd International Conference on Educational Data Mining*, R.S.J.D. BAKER, A. MERCERON AND P.I. PAVLIK JR. Eds., Pittsburgh, 311-312.
PAVLIK JR., P.I. AND ANDERSON, J.R. 2008. Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied 14*, 101–117.
PAVLIK JR., P.I., BOLSTER, T., WU, S., KOEDINGER, K.R. AND MACWHINNEY, B. 2008. Using optimally selected drill practice to train basic facts. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*, B. WOOLF, E. AIMER AND R. NKAMBOU Eds., Montreal, Canada, 593–602.
PAVLIK JR., P.I., PRESSON, N. AND HORA, D. 2008. Using the FaCT System (Fact and Concept Training System) for Classroom and Laboratory Experiments. In *Proceedings of the Inter-Science Of Learning Center Conference*, Pittsburgh, PA.
PAVLIK JR., P.I. AND WU, S. 2011. A dynamical system model of microgenetic changes in performance, efficacy, strategy use and value during vocabulary learning. In *4th International Conference on Educational Data Mining* M. PECHENIZKIY, T. CALDERS, C. CONATI, S. VENTURA, C. ROMERO AND J. STAMPER Eds., Eindhoven, the Netherlands, 277–282.
PINTRICH, P.R., SMITH, D.A., GARCIA, T. AND MCKEACHIE, W.J. 1993. Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement 53*, 801-813.
RICKARD, T.C. 1997. Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General 126*, 288-311.
RODRÂIGUEZ, M. AND SADOSKI, M. 2000. Effects of rote, context, keyword, and context/keyword methods on retention of vocabulary in EFL classrooms. *Language Learning 50*, 385-412.
THOMPSON, C.P., WENGER, S.K. AND BARTLING, C.A. 1978. How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning & Memory 4*, 210-221.
VALLACHER, R.R. AND NOWAK, A. 2007. Dynamical social psychology: Finding order in the flow of human experience. In *Social Psychology: Handbook of Basic Principles*, A.W. KRUGLANSKI AND E.T. HIGGINS Eds. Guildford Publications, New York.

VYGOTSKY, L. 1986. *Thought and Language*. MIT Press, Cambridge, Mass.
VYGOTSKY, L.S. 1978. *Mind in society*. Harvard University Press, Cambridge.
WANG, A.Y. AND THOMAS, M.H. 1995. Effects of keyword on long-term retention: Help or hindrance? *Journal of Educational Psychology 87*, 468-475.
WARD, L.M. 2002. *Dynamical cognitive science*. MIT Press, Cambridge, Mass.
WIECZYNSKI, D.M. AND BLICK, K.A. 1996. Self-referencing versus the keyword method in learning vocabulary words. *Psychological Reports 79*, 1391-1394.
WITHERSPOON, A., AZEVEDO, R., GREENE, J., MOOS, D. AND BAKER, S. 2007. The Dynamic Nature of Self-Regulatory Behavior in Self-Regulated Learning and Externally-Regulated Learning Episodes. In *Proceedings of the 13th InternationalConference on Artificial Intelligence in Education*, R. LUCKIN AND K.R. KOEDINGER Eds. IOS Press, 179-186.
WU, S.-M., YU, Y. AND ZHANG, Y. 2006. *Chinese Link: Zhongwen Tiandi, Intermediate Chinese*. Pearson Education/ Prentice Hall.