

THE EVALUATION OF ADAPTIVE SYSTEMS

Judith Masthoff

INTRODUCTION

Evaluation of their effectiveness should be an integral part of research in and development of adaptive systems. However, it is often neglected, due to many problems associated with the evaluation of adaptive systems. In this chapter, we will explore the issues, and show how we have tackled them in our research on adaptive instruction systems. Evaluations we have conducted in the past will be used as illustrations.

BACKGROUND

Empirical evaluation of adaptive systems has been rather limited. Less than one out of four articles published in *User Modeling and User Adapted Interaction* report significant empirical evaluations (Chin, 2001). Similar figures seem to hold for other journals and conferences in the field: for instance, less than 25% of the full papers presented at the 2002 *Intelligent User Interface* conference (Gill & Leake, 2002) contained an empirical evaluation. Similarly, evaluation of adaptive hypermedia systems has been very limited (Eklund & Brusilovsky, 1998). Researchers are well aware of this problem, and it led to the 2001 workshop on *Empirical Evaluations of Adaptive Systems* (Weibelzahl, Chin, & Weber, 2001) (after already having been a main focus of the *Second workshop on Adaptive Hypertext and Hypermedia*, 1998).

The research reported in the present chapter puts evaluation at centerstage. Our work focuses on adaptation in the domain of interactive instruction, producing an artificial teacher that can adapt instruction to the student (Masthoff, 1997). Our teacher is based on a community of autonomous software agents, each representing a role of a teacher, like giving feedback, navigating through course material, tailoring explanations, etc (see, Masthoff, in press, for the design and architecture of this teacher). Each agent contains a set of behaviours, specifying how the agent

should act in certain situations. These behaviors were as much as possible based on literature on human learning. A vital part of the research was to show that the artificial teacher's adaptive behavior did indeed benefit the students. Additionally, quite often we had to decide which behavior the teacher should adopt from a set of alternatives. We wanted to base these decisions on empirical evidence regarding their relative effectiveness.

ISSUES

There are many reasons why the empirical evaluation of adaptive systems has been so limited. We encountered the following when evaluating our own system.

Difficulty in attributing cause

Adaptivity is only one contributing factor to the usability of a system. So, if an adaptive system is evaluated as being 'effective' (or usable, satisfying, efficient, learnable), you cannot prove that it is the adaptation that made it effective. Similarly, if it turns out to be ineffective, it may not be the adaptation that is to blame. To overcome this problem, you can evaluate a system with adaptation to exactly the same system without adaptation (see for instance Eklund and Brusilovsky, 1998). However, this needs to be done with care, as the non-adaptive system is often not designed optimally for the task (Höök, 1998).

Adaptation is often applied to multiple aspects of a system. When evaluating a system as a whole, this makes it difficult to decide what exactly is responsible for its overall success or failure. For instance, Interbook (Eklund & Brusilovsky, 1998) provides both adaptive link annotation and direct guidance, and an experiment meant to test adaptive link annotation was spoiled by students mainly using direct guidance. Our artificial teacher (Masthoff, 1997) adapts various aspects of instruction to the student. It is quite likely that interactions occur between the various aspects, for instance, a certain way of giving feedback might work better with a certain way of giving explanations.

Adaptation itself can also be seen as consisting of multiple layers. Brusilovsky, Karagiannidis and Sampson (2001) distinguish between interaction assessment (getting a user model by interpreting user actions) and adaptation decision making (deciding the adaptation based on the model). Other authors have argued for even more layers (see, for instance, Paramythis, Totter & Stephanidis, 2001). They argue that evaluation should be conducted of each layer separately, so that it can be assessed which work and do not work.

Difficulty in finding significant results due to variance

Adaptation is normally applied to systems where the variance between users is large. After all, if the variance would be low, a system could be tailored to the user group without the need for adaptation. For instance, adaptive movie recommenders exist because users' tastes vary.

Adaptive educational systems exist because students' abilities and interests vary. Unfortunately, user variance is an enemy of statistics: the higher the variance between users, the less likely it is to get statistically significant results. This has caused problems to researchers in the past. For instance, Brusilovsky and Pesin (1998) blamed the lack of a statistically significant result in their evaluation of adaptive navigation support on the great variety in navigation styles within their subjects.

To illustrate this problem, we will briefly discuss two experiments we have preformed in the past. In the first experiment (see Masthoff, 1997), we investigated (part of) the behavior of the Practice Agent, which task is to determine the sequence of exercise items. We compared three different strategies: an often used non-adaptive strategy called Random Recycling, an existing adaptive strategy called VIP queuing, and our own adaptive strategy called Situated. We used a between subject design, with nine subjects per strategy. Unfortunately, we did not find a statistically significant effect of strategy. There was neither a significant effect for Random Recycling versus Situated or VIP queuing nor for Situated versus VIP queuing. The absence of significant effects prohibits claims regarding the relative effect of the Situated strategy. The average curves per strategy (see the left-hand graph in Figure 24) seem to indicate, nevertheless,

that on average the performance in the Situated And VIP queuing condition was better than in Random Recycling.

In the second experiment (see Masthoff, 2002), we investigated (part of) the behavior of the Navigation Agent, which task is to determine a lesson sequence. We compared three different conditions: a non-adaptive Menu condition, in which the students had to navigate themselves, an adaptive Guidance condition, in which the system selected lessons for the students, and a Mixed condition in which the students could both navigate themselves and let the system navigate for them. We used a between subject design, with fourteen subjects per strategy. Unfortunately, we did not find a statistically significant effect of condition. Nevertheless, the average results: Menu condition 33'42'', Guidance 27'12'', and Mixed 27'24'' do suggest that the Guidance and Mixed condition are better than the Menu condition. The high variance (standard deviation in the Menu condition was 14'3'') seems a likely cause of the lack in significance.

Difficulty in defining the effectiveness of adaptation

There are various ways in which the effectiveness of an adaptive system could be defined. Most often, the focus seems to be on the end result: how much information have subjects absorbed after a certain amount of time, how well do subjects perform on a test of their abilities after a certain practice time. In the case of the Practice Agent, the effectiveness of an item sequencing strategy is often measured as the average number of correct responses to a test after a certain number of practice trials. We will use the Practice Agent example to illustrate the disadvantages of defining effectiveness in this way.

In the first place, it is hard to determine after how many practice trials the test should take place. When the number of practice trials is too large, a difference in strategy effectiveness may disappear because eventually all the students learn the correct responses, regardless of the strategy used. On the other hand, if the number of practice trials is too small, the effect of the strategies on the last phase of the learning process is neglected. This problem does not only apply to the domain of interactive instruction: as adaptation often only produces benefits after the user

has been working with the system for a longer period, this period of exposure needs to be determined with care.

In the second place, even when assuming that the test takes place at the right moment in time, there is still the problem that the only aspect measured is which strategy produces the best asymptotic behaviour, i.e., with which strategy are the most difficult items learned first. Another aspect which it seems reasonable to measure is the time required to learn *most* items of the set. Of course, this poses the problem of what criterion to use for defining “most”.

In the third place, the effectiveness of a strategy may depend on the phase in the learning process. It may well be that one strategy is more effective when the student still does not know any of the items, while another strategy is more effective after a certain number of items have been learned. These kinds of effects cannot be observed with this approach.

Finally, the effectiveness of a strategy may depend on the type of subjects. It may well be that one strategy is more effective for high-ability students, while another strategy is more effective for low-ability students. Taking averages over subjects would hide this. Unexpected differences over subjects have been found before. For instance, Weber and Specht (1997) found an advantage of using adaptive link annotation for students with prior experience relevant to the subject being learned, while novices benefited more from direct guidance.

Difficulty in finding resources

The evaluation of an adaptive system requires a large number of subjects. Partly this is inherent to empirical evaluations: the power of a test (the likelihood of getting statistically significant results) increases with the number of subjects, as a rule of thumb, we like to have at least eight subjects per condition. In case of an adaptive system, you will have at least two conditions (system with and without adaptation; a within subject design is not wise, as an order effect is likely), but probably more (for the separation of concerns discussed above). The high variance within subjects (as discussed above) will also lead to needing more subjects per condition.

A second problem is that adaptation often only produces benefits after the user has been working with the system for a longer period. So, you may need subjects for a long period of time. Overall, it can make an empirical evaluation of an adaptive system both a costly and cumbersome exercise.

SOLUTIONS AND RECOMMENDATIONS

Attributing cause by separating concerns

We have decided to first evaluate each behavior of each agent separately. Only after we know the effect of individual behaviors, we will test the effect of combinations. For instance, our Practice Agent decides the order in which exercise items are presented to the student. Its simplest behavior is to present items that have been answered incorrectly more often than items that have been answered correctly (this is what we called the Situated Strategy above). On a more advanced level, it tries to distinguish between errors (the student really did not know the answer) and slips (the student knew the answer, but responded incorrectly by accident) (see Masthoff, 1997, on how this is done). In our first experiments, we only tested its simplest behavior, disabling more advanced behavior. We focussed on paired-associates learning, and left other forms of learning, such as concept learning, for later experiments (see Masthoff, 1997). We also focussed on recall (subjects had to type their responses), leaving recognition for a later experiment (subjects could select their response from a given set).

Our Navigation agent tries to adapt the lesson sequence to the goals, preknowledge, and performance of the student. In our first experiments, we ensured that the preknowledge and goals of all subjects were the same, making it possible to examine the agent's ability to adapt to the student's performance in separation (see Masthoff, 2002).

Reducing variance

For the evaluation of our interactive instruction system, we have used various, especially constructed, domains. The domains were chosen such that all subjects would have an equal pre-

knowledge (namely none), and that their content could be learned adequately within the duration of the experiment. In the Practice Agent experiment discussed above, we used the domain of learning 30 words of Japanese vocabulary (words written as pronounced). None of our Dutch subjects had any prior experience of Japanese (we asked them, and tested them on the vocabulary items at the start of the experiment). In the Navigation Agent experiment, we used the domain of Square Dancing: subjects learned how to move dancers on the screen according to the calls of the computer. Square Dancing being a typical American dance, none of our Dutch subjects possessed any knowledge (as established in a pre-experiment questionnaire). Also, the domain of Square Dancing made it possible to construct fourteen short, but highly interactive, lessons, which allowed students to complete the task within the given time, but without making the domain too small to test navigation. In both experiments, we controlled as many subject variables as possible: we used subjects of a similar age and educational background.

Defining the effectiveness of adaptation in terms of a set of learning curves

Because of the difficulties involved in having a single test score as a measure of effectiveness (as discussed above), we decided to measure learning curves in the Practice Agent experiments. The idea was that these curves would give an insight into the complete learning process, instead of concentrating on just one particular phase. We obtained these curves by alternating test and practice phases. In test phases, all 30 words were presented to the subjects in a random order. The first test was done at the start of the experiment, to establish the subjects pre-knowledge (which was indeed none). No feedback was given with respect to the correctness of the translation. In practice phases, thirty items were presented. Depending on the strategy used (the experimental condition) either all thirty words were presented in a random order (Random Recycling), or some words were presented more often at the expense of others (in VIP queuing and Situated). Feedback was given, to allow the subjects to learn.

Instead of using an average learning curve (see Figure 2), we decided to take the whole collection of learning curves (see Figure 1) into consideration in order to get an impression of

what was really going on. This implies that we did not have to define the effectiveness of a strategy in detail, but that we could use different definitions concurrently and obtain a more accurate insight into the effect of different strategies.

The Practice Agent experiment used as illustration in this chapter ended after eight test phases. We based that number on a pilot test. In subsequent experiments, we changed our method to ensure that we captured as much as possible of the learning process. We stopped the experiment once the subject's performance had reached a certain threshold (in our experiments, all items being answered correctly). We still needed a guaranteed break-off point, so that after a certain number of tests without reaching this threshold, the experiment stopped as well (to avoid over-tiring the subjects). The latter number of tests was decided in a pilot study, to limit the time to approximately an hour.

The results of the experiment are shown in Figure 1. Average results are shown in Figure 2. A MANOVA was performed on logit-transformed proportions of correct responses, with the test phase as a within-subjects repeated measures factor and the experimental condition (strategy used in the practice phases) as a between-subjects factor.

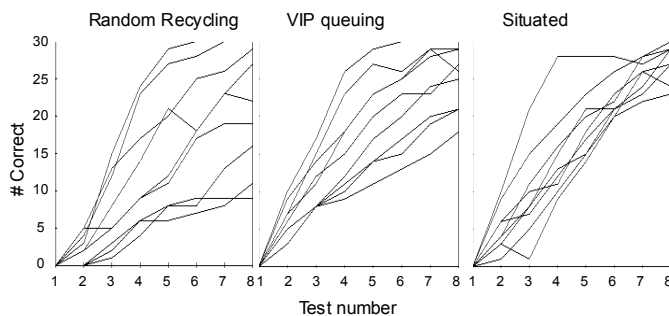


Fig. 1. Results of Experiment 1 for Random Recycling, VIP queuing, and Situated, respectively. Each line in the graphs represents a subject.

A significant interaction was found between test phase and strategy [$F(14,30)=2.5, p < .05$]. In particular, a significant interaction was found between test phase and Random Recycling versus Situated or VIP queuing [$F(7,15)=5.72, p < .01$]. Testing the effect of strategy (and the contrasts) per test phase revealed that it was only significant in the second test phase. This shows that

Random Recycling performed worse at the start of learning. A likely explanation is that both the other strategies showed some words multiple times in the first practice phase (while not showing some others), while Random Recycling showed all words exactly once. Note that this result would not have been found if we had only looked at the results of the last test.

Finding statistically significant results by splitting subjects into groups

In the Practice Agent experiment, even though there was no significant effect of strategy, Figure 1 seems to indicate an advantage of using the Situated strategy: the variance between subjects is reduced, indicating adaptation in the sense that in the Situated condition there is less difference between high and low performers (note that we would not have noticed this if we had only had the average learning curve). Therefore, a post-hoc analysis was performed in which subjects were divided into two groups per strategy: a group for the high performers and a group for the low performers. The median performance level per strategy was used as a criterion: if the number of correct responses of a subject in most test phases was above or equal to the median of that test phase, the subject was assigned to the group of high performers. On the basis of this criterion, in all strategies five subjects were assigned to the group of high performers and four to the group of low performers. The average learning curves per group and per strategy are shown in Figure 2.

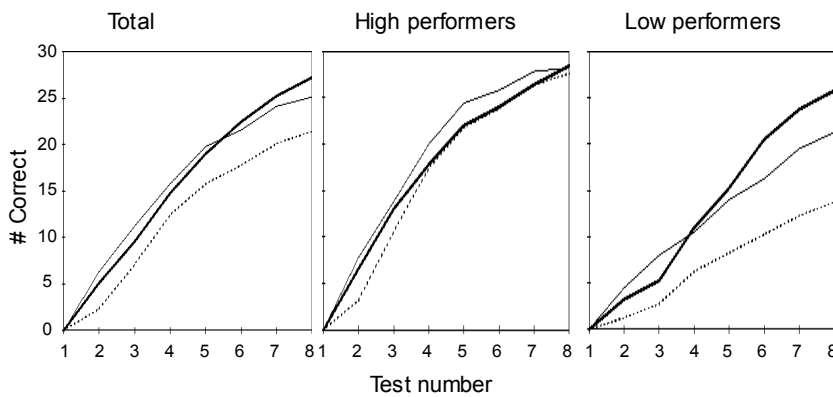


Fig. 2. Averages of the results of the Practice Agent experiment for all subjects, the high performers, and the low performers, respectively. The lines in the graphs show the average of Random Recycling (dotted line), VIP queuing (normal line), and Situated (bold line).

For each group, a MANOVA was performed on logit-transformed proportions of correct responses, with the test phase as a within-subjects factor and the experimental condition (strategy used in the practice phases) as a between-subjects factor. Both for the high and the low performers, there was a significant main effect of strategy [$F(2,1)=4.6, p < .05$; $F(2,1)=13.36, p < .01$, respectively], and a significant effect of the contrast between Random Recycling and Situated or VIP queuing [$F(1,1)=5.18, p < .05$; $F(1,1)=25.51, p < .01$, respectively]. For the high performers, the contrast between Situated and VIP queuing was significant as well [$F(1,1)=6.38, p < .05$].

These results mean that both the high and low performers benefit from using the Situated strategy and VIP queuing compared to Random Recycling. This benefit is quite large for the low performers. For the high performers, there is a small benefit of using VIP queuing compared to the Situated strategy. For the low performers, there is no significant effect of Situated versus VIP queuing, but with only four subjects per condition, the power of the test was of course very small.

We applied a similar method in the Navigation Agent experiment. The results in the Guidance condition were more homogeneous than in the Menu and Mixed conditions. The results suggested that, especially in the Menu condition, the group of subjects could be divided into two groups: a slow and a fast group. An explanation may be that in the Menu condition some subjects merely follow the sequence of the menu, while others are much better at deciding which lessons are relevant. The heterogeneous results in the Mixed condition may be explained by some subjects being more reluctant to use guidance than others. Both these explanations are confirmed by the navigation paths, and by the time spent per lesson.

A post-hoc analysis was performed in which the subjects of both the Menu and Mixed conditions were divided into two groups: one group with subjects who studied at least two superfluous lessons (lessons not contributing to the overall goal of learning the Slide Through move), and

one with the remaining subjects. We will call the first group the poor navigators and the second group the good navigators. In the Menu condition, nine subjects belonged to the poor navigators, and five to the good navigators. In the Mixed condition, four subjects belonged to the poor navigators, and ten to the good. Figure 3 shows the results per subgroup. Three contrasts were performed. Significant effects of condition for the contrasts between the poor and good navigators [$F(1,1)=22.05p<.0033$], and the poor navigators and the Guidance condition [$F(1,1)=11.31p<.00033$] indicate that the poor navigators need more time than the good navigators and the subjects in the Guidance condition. This implies an advantage of using guidance for students who are unable to monitor their own learning process. As nine of the fourteen subjects of the Menu condition are poor navigators, determining a good path through the course material seems a difficult task, therefore guidance seems necessary.

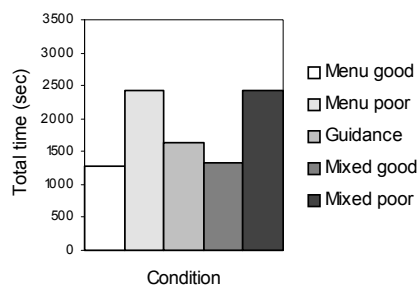


Fig. 3. Average total times per group for the Navigation Agent experiment.

Reducing the amount of resources needed by using experts instead of subjects

The dream of every system developer and researcher is to be able to evaluate their systems without having to use human subjects. Methods have been designed that allow this to a certain degree. Cognitive Walkthrough and Heuristic Evaluation are popular methods for evaluating a user interface design without using subjects (Nielsen & Mack, 1994). Both employ usability experts instead. In the cognitive walkthrough the experts pretend to be users, very naïve ones that are new to the system. They follow the correct action sequence for a user's task, and for each step determine whether it would be a success or a failure, using four trigger questions. In the heuristic evaluation, experts use a set of guidelines (heuristics) about what a good user interface should be

like, and judge the design against these guidelines. Some guidelines have a special meaning in the case of an adaptive system. For instance, the guideline about system status "the user should always know the state of the system", can be interpreted as asking for transparency: making sure the user understands why adaptation occurs, how the system got to its existing (adapted) state, perhaps even making the user model inspectable, and understandable. The guideline about user control can be interpreted as users should be able to both undo system adaptations and disable automatic adaptations altogether.

Both methods provide a good basis for improving the design. We tend to use them to make sure there are no obvious user interface problems in our systems before an empirical evaluation.

However, it is always recommended to complement this kind of evaluation with a user test. This is mainly because it is very hard for an expert to pretend to be a naïve user, and it is difficult to predict all problems users will have. Also, it is difficult to use these methods to predict the effectiveness of adaptation strategies: though they can predict many usability problems, they cannot really predict how fast, for instance, a user will learn.

Reducing the amount of resources needed by using models instead of subjects

Another way to remove the need for human subjects would be to replace them with an executable computer model. The model should be able to perform the user's task, and act like real users would do, experiencing similar problems. It has been one of the main aims of cognitive science to construct and evaluate models of human behavior controlled by human cognitive processes. In the domain of learning, various models exist, including models of paired associates learning (e.g., Bower, 1961) and concept learning (e.g. Kruschke, 1992).

Examples of models

Markov models are simple models which are mostly used in cognitive psychology to model the learning of paired associates. A Markov learning model can be characterized by a set of knowledge states, a transition matrix which indicates the probability of a transition from one

knowledge state into another, and a response function which maps knowledge states onto probabilities of responses. The most important property of a Markov model is that the knowledge state in which a student is at a certain time $t+1$ only depends on the knowledge state at time t . The *All-or-None model* of Bower (1961) is a very simple Markov model with two states, say G (guessing) and M (mastered) (see Figure 4). An item can be either completely mastered, in which case the student will always give the correct response, or it can be completely unknown, in which case the student can only guess the correct response, say with likelihood g . Once an item is mastered, it will always be known, so there is no forgetting. An item has a certain likelihood, say α , to be learned each time it is presented.

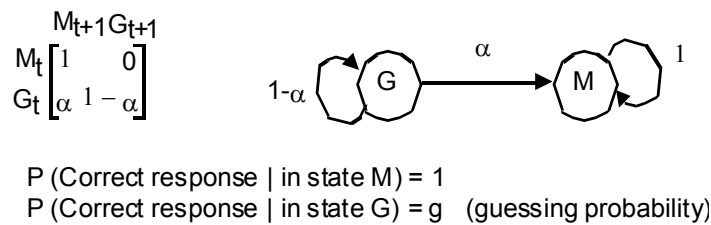


Fig. 4. Transition matrix, corresponding diagram, and response function for the All-or-none model.

To model differences between students, a different value of α can be used for each student: the higher the α , the faster the student learns. To model differences between item difficulties, a different value of α can be used for each item: the higher the α , the easier the item. To model differences in the interaction between students and items, different values of α can be used for each combination of student and item: the higher the α , the easier the item is for that particular student.

Markov models can be made more sophisticated by adding more states and associating a parameter with each state transition, or by making the response function more complex.

The assumption that learning may be partial, in the sense that stages may be distinguished in the learning process, would result in the incorporation of more states in the model: one for each stage. An extreme version of this is the *Linear model* (Atkinson, Bower, & Crothers, 1965) which can be viewed as a Markov model with an infinite number of states, and is shown in Figure 5. In this model, the probability of a state transition is 1. The assumption underlying this model is that each presentation of an item reduces the error probability by a constant factor, say α .

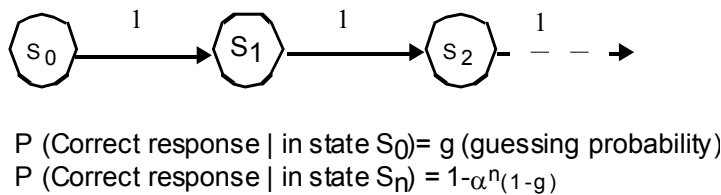


Fig. 5. Diagram and response function for the Linear model.

One advantage of using of Markov models is that they are mathematically well understood. However, they are black box models in the sense that item characteristics can only be taken into account indirectly: through parameter values. Interactions in the learning of different items are difficult to incorporate into these models.

Connectionist models such as ALCOVE (Kruschke, 1992) have become very popular, especially for concept learning. An advantage of using connectionist models is that they can give more insight into the learning process, for instance as regards the interaction in learning a set of items. However, item characteristics have to be defined beforehand, and mathematically they are more complex than, for example, Markov models.

Before models can really be used to replace humane subjects, they have to be validated. One way to validate them is to use the models to predict the outcome of experiments, and then compare their predictions with the real results as produced by human subjects. Another way is to investigate how well results from human subjects can be reproduced by the models. We have used both these approaches in our evaluation of the Practice Agent.

Using models to predict the outcome of experiments

We have used both the All-or-None model and the Linear model (and some others) to predict the outcome of the experiment with the Practice Agent that was described above. The results of the simulations of students by the All-or-None model are shown in Figure 6. The All-or-None model predicted a clear advantage of using the Situated strategy compared to the other strategies: all the learning curves are higher in the case of the Situated strategy.

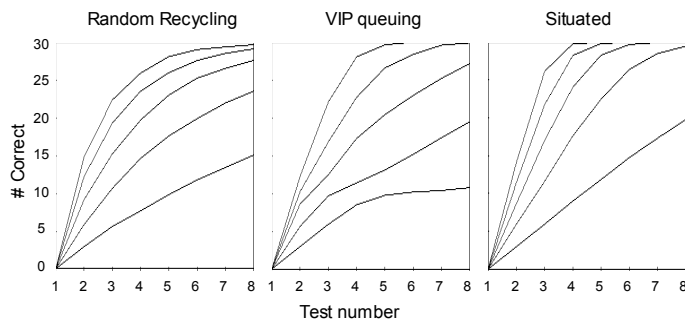


Fig. 6. Predictions of the All-or-None model for Random Recycling, VIP queuing, and Situated, respectively. Each line in the graphs represents the average over 100 runs of the model with parameter α (from bottom to top) 0.1, 0.2, 0.3, 0.4, and 0.5, respectively

The model also predicted VIP queuing to be more effective than Random Recycling for very good students: the increase in the two highest learning curves for VIP queuing lasts longer than for Random Recycling and therefore the curves end higher. However, it predicted VIP queuing to perform rather poorly for weak students: the two lowest learning curves for VIP queuing end lower than those for Random Recycling and the lowest learning curve especially hardly increases at all in the end. A possible explanation for this is a kind of blockage effect (see Masthoff, 1997 for an explanation).

The results of the simulations of students by the Linear model are shown in Figure 7.

Mathematically, Random Recycling is the optimal strategy when the Linear model is correct.

Indeed, the simulations show an advantage of using Random Recycling. However, the results for the Situated strategy approximate the results of Random Recycling: the difference is rather small.

VIP queuing performs worse than Situated, even for the best students. This may again be explained by a blockage effect.

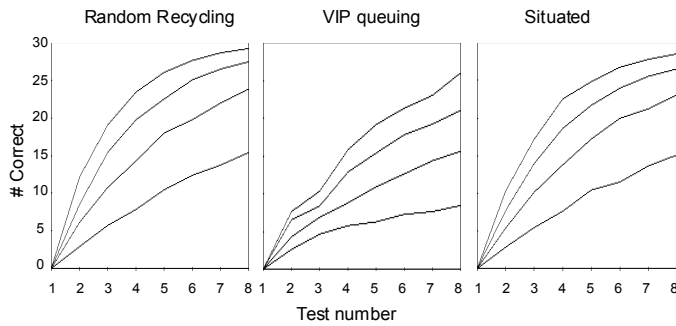


Fig. 7. Predictions of the Linear model for Random Recycling, VIP queuing, and Situated, respectively. Each line in the graphs represents the average over 100 runs of the model with parameter α (from top to bottom) 0.6, 0.7, 0.8, and 0.9, respectively.

We expected that the blockage effect would become worse when not all items were equally difficult: more difficult items may block the learning of easier items. It also seemed likely that the advantage of using the Situated strategy would become even larger when not all items were equally difficult. To check the correctness of these hypotheses, simulations were done with the All-or-None model, with a set of items of varying difficulties. We have assumed that parameter α of the model (representing the transition probability from the initial to the mastered state) is the multiplication of a parameter β which represents a student's learning capability (the higher the value of β , the faster the student learns) and a parameter γ which represents an item's difficulty (the higher the value of γ , the easier the item). Parameter β has been varied between 0.01 and 0.1. Parameter γ has been kept at 10 for 10 of the 30 items, and at 1 for the other 20 items. Hence, one third of the items are 10 times as easy to learn as the other items. The results of the simulations are shown in Figure 8.

Two observations can be made. Firstly, our hypothesis regarding the blocking effect seems correct. All the learning curves in the case of VIP queuing are very low, approximately as low as, or even lower than, the lowest learning curve in Figure 6. Secondly, the Situated strategy is more

effective than Random Recycling, especially for the higher learning curves. However, the effect is not as great as we might have expected. It can be observed that Random Recycling scores better in the beginning, at the second test phase. The learning curves start steeper. This can be explained by the fact that in Random Recycling all items, including all easy items, are presented to the student in the first practice phase. Hence, the student has an early opportunity to learn these easy items. On the other hand, in the Situated strategy it is possible that some of the easy items have not yet been presented.

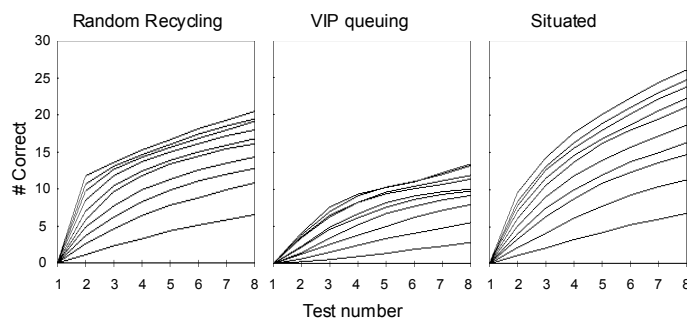


Fig. 8. Predictions of the All-or-None model in the case of varying item difficulties for Random Recycling, VIP queuing, and Situated, respectively. Each line in the graphs represents the average over 100 runs of the model with parameter β (from bottom to top) from 0.01 to 0.1.

Fit of the models

The fits were obtained by using the same procedure as in the case of the predictions. However, in case of the fits, the parameters of the models were varied while calculating which parameter setting produced the best fit to the experimental data. Instead of fitting the model to the average of the subjects in a certain experimental condition, we opted to fit all subjects individually. In that way, the model also has to explain the variance between the subjects. We used a least squares measure, in which the fitting process minimises the sum over test phases of the squares of the difference between the number of correct responses of a subject and the number of correct responses of the model (denoted lsq). Pearson correlation coefficients were determined between the experimental data and the fit per strategy, and between the data and the fit per strategy per test phase.

Both the All-or-None model and the Linear model were fitted by varying the value of parameter α . For the All-or-None model, the mean and standard deviations of the lsq were $m=41.27$, $sd=33.14$ for Random Recycling, $m=12.52$, $sd=6.81$ for VIP queuing, and $m=19.39$, $sd=12.26$ for the Situated strategy. For the Linear model, the mean and standard deviations of the lsq were $m=179.21$, $sd=44.48$ for Random Recycling, $m=8.89$, $sd=4.82$ for VIP queuing, and $m=97.67$, $sd=60.01$ for the Situated strategy. The learning curves corresponding to the fits are shown in Figure 9 for the All-or-None model, and Figure 10 for the Linear model.

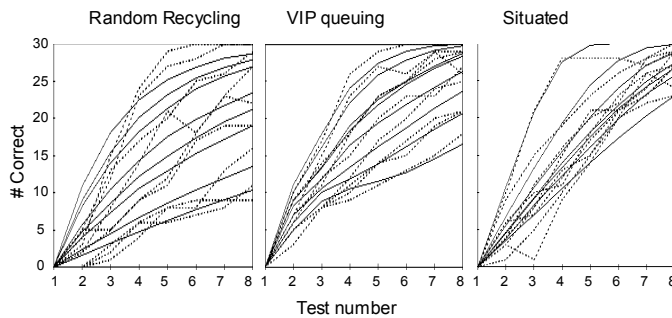


Fig. 9. Results of the fit of the All-or-None model on the data of Experiment 1 for Random Recycling, VIP queuing, and Situated, respectively. Each solid line in the graphs represents a fitted subject, each dashed line represents a real subject.

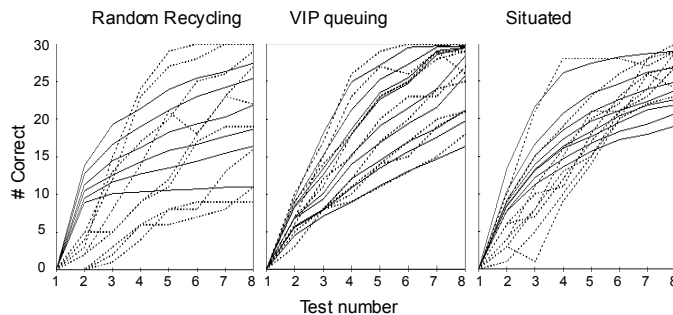


Fig. 10. Results of the fit of the Linear model on the data of Experiment 1 for Random Recycling, VIP queuing, and Situated, respectively. Each solid line in the graphs represents a fitted subject, each dashed line represents a real subject.

For the All-or-None model, both the correlation coefficients and the near-coincidence of the learning curves of the model and those of the real subjects (see Figure 9) indicate that the model fits the data very well. Only the beginning of the curves, particularly in the second test phase, is fitted relatively poorly, especially in the case of Random Recycling (correlation coefficient of .689). The learning curves of the real subjects start less steeply than those of the model. However, even if the relatively bad fit of the beginning of the curves is neglected, the good fit does not imply that the model is correct. There is definitely an effect of strategy on the α distribution. According to this fit, the subjects in the VIP-queuing condition were relatively good (high α values) and those in the Situated condition were relatively bad (low α values). This is not very likely, given the random assignment of subjects to conditions.

For the Linear model, both the correlation coefficients and the obvious difference between the learning curves of the model and those of the real subjects for both Random Recycling and the Situated strategy (see Figure 10) indicate that the model fits the data rather poorly. The data of the Situated strategy, especially, are fitted very poorly. With this model, the second test phase of Random Recycling also shows an relatively poor fit, with a much steeper increase for the model than for the real subjects. Moreover, the α distribution is unequal for the different strategies.

The usefulness of the models

A comparison of the fits and predictions of the models with the experimental data showed two major shortcomings of the models. In the first place, all the models discussed produced steeper starts of the learning curves than the real subjects. In the second place, all the models largely underestimated the effect of VIP queuing. This suggests that they lack a mechanism which explains the positive features of VIP queuing. Such a feature may be that the cognitive load of the student is reduced by not presenting all the items at once, but only presenting an item when most of the items previously presented have been learned. Our Situated strategy clearly lacked such a feature. So, even though the models obviously are not good enough yet to replace the use

of real subjects, the differences between the predictions and the experimental data actually help to understand more in detail what happened and how the strategies can be improved.

CONCLUSION

Evaluating adaptive systems is not easy. However, it is vital to ensure scientific progress, and to provide convincing arguments that adaptation really does help. In this chapter, we have indicated some ways forward. In particular, we believe that studying individual learning curves gives more insight in experimental data than average, one-test results. Splitting subjects into groups can sometimes produce more insight and better results. The use of executable models, that can simulate human behavior, can both make it less costly to run experiments, and provide more understanding of what causes certain experimental results. It remains a major challenge to construct executable models that accurately predict human behavior. However, we have shown how even the weaknesses of the models can help to clarify the issues.

REFERENCES

- Atkinson, R.C., Bower, G.H., & Crothers, E.J. (1965). *An introduction to mathematical learning theory*. New York: John Wiley and Sons
- Bower, G.H. (1961). Application of a model to paired-associate learning. *Psychometrika*, 26, 255-280.
- Brusilovsky, P., Karagiannidis, C., and Sampson, D. (2001). The benefits of layered evaluation of adaptive applications and services. In S. Weibelzahl, D. Chin and G.Weber (Eds.), *Empirical evaluation of adaptive systems: Proceedings of workshop at the Eighth International Conference on User Modeling, Freiburg*.
- Brusilovsky, P. and Pesin L. (1998). Adaptive navigation support in educational hypermedia: an evaluation of the ISIS-tutor. *Journal of computing and information technology*.
- Chin, D. (2001). Empirical evaluation of user models and user-adapted systems. *User modelling and user-adapted interaction*, 11, pp 181-194.

- Eklund, J. and Brusilovsky, P. (1998). The value of adaptivity in hypermedia learning environments: A short review of empirical evidence. In proceedings of the Adaptive Hypermedia workshop.
- Gill, Y, and Leake, D.B. (Eds.) (2002). IUI 02: proceedings of the 2002 International conference on intelligent user interfaces. San Francisco: ACM press.
- Höök, K. (1998) Evaluating the utility and usability of an adaptive hypermedia system. *Journal of Knowledge-Based Systems*, 10 (5).
- Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.
- Masthoff, J. (1997). An agent-based interactive instruction system. Ph.D. thesis Eindhoven University of Technology.
- Masthoff, J. (2002). Design and evaluation of a navigation agent with a mixed locus of control. In S.A. Cerri, G. Gouardères, F. Paraguaçu (Eds.), *Intelligent Tutoring Systems: Proceedings ITS 2002*. Berlin: Springer Verlag.
- Masthoff, J. (in press). APPEAL: A domain-independent artificial teacher. *International journal of continuing engineering education and lifelong learning*, 12. Special issue on intelligent agents.
- Nielsen, J., & Mack, R.L. (Eds.) (1994). *Usability Inspection Methods*. John Wiley & Sons, New York.
- Paramythis, A, Totter, A., & Stephanidis, C. (2001). A modular approach to the evaluation of adaptive user interfaces. In S. Weibelzahl, D. Chin and G. Weber (Eds.), *Empirical evaluation of adaptive systems: Proceedings of workshop at the Eighth International Conference on User Modeling*, Freiburg.
- Weber, G., Specht, M. (1997). User modeling and adaptive navigation support in WWW-based tutoring systems. In Proceedings of the 6th international conference on user modeling.

Weibelzahl, S., Chin, D., & Weber, G. (Eds.). (2001). Empirical Evaluation of Adaptive Systems. Proceedings of workshop at the Eighth International Conference on User Modeling, UM2001. Freiburg.

Wiebelzahl, S., and Weber, G. (2002). Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. *Kunstliche Intelligenz*, 3. As accessed on [http://home_ph_freiburg.de/wibelza/](http://home.ph.freiburg.de/wibelza/) 24 April 2002