Can we help students with a high initial competency?

Carolyn P. Rosé and **Kurt VanLehn** and **Pamela Jordan**

LRDC, University of Pittsburgh
Pittsburgh, PA 15260 USA
rosecp@pitt.edu

Abstract

It is a well known phenomenon that students with high initial pretest scores demonstrate lower learning gains since there is less room for improvement. In this paper we explore a related issue, that of selecting appropriate interventions for students who start out with a high initial competency. We present a mathematical model that makes it possible to project success rate at selecting appropriate interventions based on accuracy at analyzing student performance at a task. This model demonstrates that for high initial competency students, selecting interventions based on any sort of isolated analysis of student performance is doomed to failure even with very high accuracy analysis. We explore an alternative approach to selecting interventions within the context of the WHY2 conceptual physics tutoring system (VanLehn et al., 2002) based on prior probabilities of student need. Our analysis demonstrates that within this domain, individual differences between high competency students makes this type of approach equally unsuccessful. We conclude by discussing some alternative approaches to solving this problem, which we are still investigating.

Keywords: knowledge construction dialogues, selecting interventions

1 Introduction

It is a well known phenomenon observed in all types of learning studies that students with high initial pretest scores demonstrate lower learning gains since there is less room for improvement. In this paper we explore a related issue, that of selecting appropriate interventions for students who start out with a high initial competency. We use as our example the WHY2 conceptual physics tutoring system (VanLehn et al., 2002). The goal of the WHY2 system is to coach students as they explain physics mental models in natural language in response to short essay questions such as, "Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain." The WHY2 system has at its disposal a library of knowledge construction dialogues (KCDs), i.e., interactive directed lines of reasoning, each of which is designed either to elicit a specific idea (i.e., an elicitation KCD) or to remediate a specific misconception (i.e., a remediation KCD).

When students interact with WHY2, they are first presented with an essay question. After reading the essay question, the student types in an initial essay. The system then analyzes the student's essay in order to detect the presence of misconceptions and required concepts. The system then uses KCDs both for coaching students to insert missing required concepts (elicitation KCDs) and to remove the expression of misconceptions (remediation KCDs). When a student's essay is analysed, the system constructs a list of missing required points and misconceptions. For each of these items, the student will receive an elicitation KCD for every point analysed as missing and a remediation KCD for every misconception analysed as present.

As the system interacts with the student, it keeps track of which interventions have already been given, and thus can avoid repeatedly offering the same intervention. This contextual information about which

interventions have been offered can be thought of, then, as a sort of primitive user model that aids in future decisions about which of the available interventions it should offer. However, for the initial essay that students type in, the system makes its decision about what intervention to offer purely based on its analysis of the student's essay. Intuitively, helping students with a low initial competency is easier than for those with a high initial competency since almost any intervention selected is likely to be appropriate since so many are needed. Equally intuitive is the idea that high analysis accuracy is required for selecting interventions for high initial competency students since they need only a very few interventions. Thus, the likelihood is high that a haphazardly selected intervention will not be appropriate, and conversely that the few interventions that are required will not be selected. What is less intuitive is how dire the situation actually is once this phenomenon is quantified, and the conclusion that selecting interventions based solely on an isolated analysis of student performance is doomed to failure with high initial competency students. One possible solution is to use prior probabilities of student need to influence the selection of interventions. Our preliminary investigations of this approach within the context of WHY2 have demonstrated that the high degree of individual differences in our domain between high competency students make some sort of student specific student modeling necessary to make this type of approach successful.

We begin by presenting our model for projecting success at selecting appropriate interventions based on accuracy of analysis of student performance. We then describe our preliminary investigations into using prior probability of student need to influence the selection of interventions in WHY2. We conclude with some current directions.

2 Selecting Appropriate KCDs

Figure 1: **This Figure summarizes our model for predicting KCD precision, recall, and false alarm rate from analysis precision, recall, and false alarm rate. Note that this model only applies to the selection of elicitation KCDs.**

	Point in Essay	Point Not in Essay
Point Identified	A	B
Point Not Identified	C	D

Analysis Precision	$= A/(A+B)$
Analysis Recall	$= A/(A+C)$
Analysis False Alarm Rate	$= B/(B+D)$
KCD precision	$= D/(C+D)$
KCD Recall	$= D/(B+D)$
KCD False Alarm Rate	$= C/(A+C)$

In order to build an effective system, it is important both to give students the interventions that they do need, and to avoid giving them extraneous interventions that they do not need. Neglecting to give a student an intervention that is needed means losing an opportunity to teach that student something that student needs to know. Giving an intervention that a student does not need means wasting a student's time, possibly distracting that student from what that student really needs to learn, and likely annoying or even confusing that student. Thus, we would like to build a system with a high interventions selection recall and low selection false alarm rate, where we define selection recall as the percentage of interventions that a student needs that the system gives. And selection false alarm rate as the percentage of interventions that the student does not need that the system gives.

Nevertheless, analyzing student performance, which in WHY2 is an essay, but in another system may be an equation or a proof, etc., is most naturally evaluated separately from that of selecting interventions. For example, in the case of WHY2, analysis of student essays is a computational linguistics problem, and performance on this task is most naturally conceptualized as a text classification problem and measured

in terms of analysis precision, recall, and false alarm rate over a corpus of student essays. Analysis precision is the percentage of required points and misconceptions identified in the student essays that were actually present in those essays. Note that this is undefined in the case that no required points are identified. Related to this notion is analysis false alarm rate, which is the percentage of required points not present in the essay that were incorrectly identified by the system. Analysis recall is the percentage of misconceptions and required points present in student essays that were actually identified by the system. Note that this is undefined whenever there are no required points present in a student essay. Naturally, a system that is good at accurately identifying required points and misconceptions in student essays will also be good at selecting appropriate KCDs to engage students in. However, our mathematical model demonstrates that the relationship between analysis precision, recall, and false alarm rate and KCD precision, recall, and false alarm rate varies widely depending upon the quality of student essays.

Figure 1 presents some equations that describe our mathematical model. Note that this model only applies to the selection of elicitation KCDs. In our domain, missing information from essays is a much more prevalent problem than incorrect information indicating a misconception that we are prepared to handle with an available remediation KCD. As the equations in Figure 1 demonstrate, we define Recall for analysis as the number of required points that WHY2 correctly identifies as present in a student essay divided by the total number of required points actually present in the essay. Precision is the number of required points correctly identified divided by the total number of required points that WHY2 identified, correctly or incorrectly. False alarm rate is computed by dividing the number of required points identified but missing in the essay by the total number of required points missing from the essay. Note that analysis precision, recall, and false alarm rate are a byproduct of the analysis approach we have selected. Different NLU techniques achieve different levels of precision, recall, and false alarm rate. And we are continuing to experiment with different approaches. However, once an analysis approach is selected, we treat analysis precision, recall, and false alarm rate, which measure our accuracy at analysing student essays, as given. On the other hand, you will notice that KCD precision, recall, and false alarm, which measure our accuracy at selecting elicitation KCDs, is a function of analysis accuracy and essay quality. It is also possible to show that, holding KCD accuracy constant, that analysis accuracy varies with essay quality. But in practice, it makes more sense to consider analysis accuracy as inherent in the NLU approach, and KCD accuracy is derived from that. Note that we define essay quality here as the percentage of required points that are included in the student's essay. It does not take into consideration any expressions of misconceptions or wrong information also present.

Now let's consider how KCD precision, recall, and false alarm rate are related to analysis precision, recall, and false alarm rate as well as essay quality. As you see from the equations in Figure 1, KCD precision is the number of KCDs correctly given divided by the total number of KCDs given. KCDs are given whenever a required point is not identified. Thus, when analysis recall is low, a lot of points that are present in the student's essay will not be identified. Thus, the corresponding elicitation KCDs will be incorrectly given. A needed KCD is not given whenever a point is incorrectly identified in a

Figure 2: **This Table illustrates how KCD precision and recall vary with essay quality, keeping 0.90 analysis precision and 0.90 analysis recall.**

Essay Quality	KCD Precision	KCD Recall
0.10	0.99	0.99
0.20	0.98	0.98
0.30	0.96	0.96
0.40	0.93	0.93
0.50	0.90	0.90
0.60	0.85	0.85
0.70	0.77	0.77
0.80	0.60	0.60
0.90	0.10	0.10

student's essay. Thus, when analysis precision is low, many of the points that are identified are not actually present in the student's essay. Thus, many of the KCDs that the student needs will not be given. KCD false alarm rate is the percentage of KCDs that should not be given but are. Thus, KCD false alarm rate is 1 - analysis recall.

KCD precision and recall are influenced both by analysis precision and recall as well as essay quality. To illustrate this point, let is consider the equations in Figure 1. Notice that A represents the percentage of time that a point is in the essay and identified as in the essay. Notice further that B represents the percentage of time that a point is not in the essay but identified as in the essay anyway. C represents the percentage of time that a case is in the essay and not identified as such. Finally, D represents the percentage of time that a point is missing from the essay and not identified as being in the essay.

When essay quality is high, D will be low. If essay quality is 70%, D can be no greater than 30%. Remember that KCD precision is $D/(C+D)$. If analysis recall is perfect, then C will be 0%, so KCD precision will be 100%. But if analysis recall is less than perfect, even a little bit less, then KCD precision goes down very fast. Since D is small, even if C is equally small, KCD precision is down to 50% already. KCD recall is determined by the relationship between B, the points not in the essay that were identified, and D, the points not in the essay that were not identified, in particular, $D/(B+D)$. Thus, if analysis precision is perfect, then B will be 0%, so KCD recall will be 100%. But if it is not perfect, then it will bring the KCD recall down fast, again because D is necessarily small.

The opposite is the case when essay quality is very low, lets say 20%. In this case, A and C must both be low. Thus, even if analysis recall is 0%, C will be no more than 20%. Let's say that analysis recall and precision are around 86%. Whenever essay quality is low and analysis precision is reasonably high, B and C must both be small in comparison with D. Analysis precision and recall of 86% would have been disasterous for a high quality essay as we saw above. However, in this case, it would mean that A is 17%, B is 3%, C is 3%, and D is 77%. Thus, KCD precision and recall are both 96%. Therefore, the situation is quite different when essay quality is low. If analysis accuracy is reasonably high, KCD assignment accuracy will also be high.

Therefore, from this mathematical model we can see that as essay quality increases, it becomes much more difficult to do a good job at selecting appropriate KCDs for students. In fact, selecting appropriate KCDs for students with essay qualities of 70% or higher may well be completely out of our reach. In particular, even if analysis precision and recall are almost perfect, specifically at 90%, KCD precision, recall, and false alarm rate become unsatisfyingly low once essay quality is 70See Figure 2. Figure 3 shows the projection of KCD precision and recall from our current best analysis performance.

Thus, when we depend upon our analysis of student's essays to tell us how to select interventions for students, we may be at a loss with respect to helping high end students who produce high quality essays. One might wonder whether it is necessary to worry about the high end students since it is often said that good students will learn even with a bad teacher. Nevertheless, no matter how good the students are, if we fail to present them with instruction on the topics that they lack, they will not have the opportunity

Figure 3: **This Table illustrates how KCD precision and recall vary with essay quality, keeping 0.79 analysis precision and 0.80 analysis recall, which is our current best performance with** WHY2.

Essay Quality	KCD Precision	KCD Recall
0.10	0.98	0.98
0.20	0.95	0.95
0.30	0.91	0.91
0.40	0.87	0.86
0.50	0.80	0.79
0.60	0.69	0.68
0.70	0.52	0.50
0.80	0.16	0.15
0.90	0.00	0.00

to learn those topics. Additionally, we have evidence in the form of expressions of frustration in our log files that students are unhappy when they receive help on many topics that they do not need help on. However, we do not have a definitive answer on what is the minimum acceptable level of KCD precision, recall, and false alarm rate. This is still under investigation.

Interestingly, a similar model has been proposed in the medical field[1]. In that field, it has been determined based on this similar model that the Positive Predictive Value of a diagnostic test depends both on what is called the Specificity as well as the Prevalence of the disease in the population. What is called Sensitivity in that field is analogous to our analysis recall. What they call Specificity is analogous to our KCD precision. What they call Positive Predictive Value is analogous to our analysis precision. And their Negative predictive value is analogous to our KCD precision. Prevalence of disease could be thought of as analogous to our essay quality, or percentage of required points that are present in an essay. It is possible to verify that these pairs represent analogous quantities since both are defined in terms of true positives, true negatives, false positives, and false negatives. In their model, Positive predictive value equals (Sensitivity times Prevalence) divided by ((Sensitivity times Prevalence) + (1 - Specificity)(1 - Prevalence)). Although they are solving for a different quantity than what we are primarily interested in, their model is interesting in that the relationship between the four quantities in their model is the same as that in our analogous model. Thus, the problem of high prevalence (or analogously, high essay quality) is a common problem, and we may learn something from their solutions to their analogous problem. In the medical field, the problem is addressed by using multiple diagnostic tests. Thus, perhaps an analogous approach would prove beneficial here as well.

3 Using Prior Probabilities of Student Need

An evaluation of the WHY2 system is currently underway with undergraduate students having already taken one college level physics course. We examined 95 student essays collected from 22 students interacting up until now with either one of the two different versions of the WHY2 system (WHY-Atlas and WHY-AutoTutor) to determine how often students type high quality essays. Specifically we looked at essays for the pumpkin problem introduced above. We determined that 21% of student essays (i.e., 20 out of 95 essays) are in the high competency range (i.e., missing 2 or fewer of the 6 required points for the pumpkin problem), however, only 10% of these essays (2 essays) are initial essays. Thus, for the majority of high quality student essays, both versions of WHY2 have access both the analysis of the essay as well as the dialogue history (i.e., record of which interventions have already been given). Thus, the problem is not as dire in practice as it would be if the system did not have access to the dialogue history. However, it is instructive to note that of the 20 high quality essays, 75% of them are at the 67% essay quality level (i.e., missing 2 out of the 6 required points). Thus, very few students are getting past high-mediocre performance at constructing quality explanations. Further analysis is required to determine with high confidence whether this is primarily due to lack of success at selecting appropriate interventions or to some other problem.

To address the problem of selecting an appropriate intervention for a high quality essay, we explored the possibility of selecting based on prior probability of student need rather than on an analysis of student essay quality. Our preliminary analysis demonstrates, however, that this is unlikely to be a solution to the problem. We first checked the probability of an essay including each of the required points over the whole corpus of 95 student essays. We then compared these probabilities with those of in the 20 high quality essays. For ease of reference we will refer to the points as A, B, C, E, J, and K respectively. Point K was the most likely to be included in a student essay. It appeared in 79% of the essays overall and 100% of the high quality essays. Point K occurred in 44% of essays overall and 85% of high quality essays. Point A occurred in 39B occurred in 27% of essays overall and 50% of high quality essays. Point J occur5ed in 28% of essays overall and 75% of high quality essay. Finally, point E occurred in 18% of essays overall and 50% of high quality essays. Note that the ranking by probability of occurrence for points is not the same for high quality essays as it is for essays overall. Thus, the first caveat we encountered in our analysis was that prior probabilities should be computed over just high quality essays, and not over the entire set of essays. This means that in order to get reliable statistics,

[1] See www.ipathology.com/Disease%20Prevalence%20Notes.htm and www.epidemiolog.net/studymat/PredictiveValue.xls.

since high quality essays occur only 20% of the time, we need a tremendous amount of data in order to collect a large enough sample of high quality essays to base the statistics on.

Even if we assume that the current set of 20 high quality essays is sufficient to obtain a probability distribution that is good enough for our purposes, the numbers we get do not seem to help us do a better job at selecting appropriate interventions, even for that same set of essays, because of the high degree of individual differences in student performance on high quality essays. The three points least likely to occur in high quality essays are E, B, and J. However, if the corresponding elicitation KCDs for all three of these were given to students who wrote those high quality essays, roughly 2/3 of the KCDs that would be given to students would be inappropriate. This is because the two points on average that are missing from these high quality essays varies widely over the set of high quality essays. If only the top 2 rather than the top 3 interventions were given instead, about half of the KCDs given would be inappropriate, and students would also be missing about half of the KCDs that they should get. Considering that these high quality essays are primarily at the 67% quality level, this prior probability based approach puts us in the same bad position with respect to selecting appropriate interventions than even our current best analysis approach (see Figure 3). Some combined approach, taking both the analysis and prior probabilities might prove to be the most effective approach. A more likely approach to be successful would be to compute some student specific model of prior probability of needing a specific intervention, possibly based on the student's pretest performance. Again, further analysis is required to determine whether this is the case.

Another possibility would be to take a different approach altogether for high quality essays. Perhaps we could take the production of a high quality essay as an indication that the student might learn better with a slightly harder problem. If the answer to the slightly harder problem required much of the same reasoning on a deep level as the easier problem, then the harder problem could be used to teach the same concepts that the student showed evidence of needing instruction on with the easier problem. If we take a very conservative approach to essay analysis, such as using LSA with a high threshold, we can achieve a high analysis precision with an accompanying low recall. If we took such an approach, we would rarely miss opportunities to teach a student something that was needed, but we would teach on lots of things that were not needed. However, if we teach these same concepts in the context of a harder problem, it is less likely that the instruction will seem superfluous. Thus, such an analysis approach could work well.

As an illustration, let us consider a harder version of the pumpkin problem. Remember that the original pumpkin problem is as follows: "Suppose you are running in a straight line at constant speed. You throw a pumpkin straight up. Where will it land? Explain." In the original pumpkin problem, the man throws the pumpkin straight up. But consider what happens when the man exerts both a vertical and a horizontal force on the pumpkin during the toss. In the original problem, the man's horizontal velocity remains equal to the pumpkin's horizontal velocity. Thus, their horizontal displacements from the point of release will always be equal. And when the pumpkin lands, it will land in the man's hands. But in the harder case, the student must reason that during the toss, the pumpkin will accelerate in the horizontal direction. Thus, although both the man and the pumpkin will move at a constant horizontal velocity after the toss, the pumpkin's velocity will be greater. Thus, it's horizontal displacement from the origin will always be greater than that of the man. And it will land in front of the man.

If the student produces a high quality essay in response to the easy version of the problem, the system could reanalyze the student's essay with a very stringent, conservative approach. Then, we could tutor on all of the required points analyzed as missing, but in the context of the harder problem. For example, the system could say, "Since your essay looks very good, let me give you a harder scenario..." Then, for example, if the analysis of the essay for the easy problem indicated that the student did not articulate the fact that the horizontal displacement of the man and pumpkin would always be the same, the system could ask, "In this new scenario, what do you think will be the relationship between the man's horizontal displacement and that of the pumpkin at all times after the release?" If the student can answer this question correctly, the system can assume that the student did not need to be tutored on the corresponding topic in the easier scenario since the same reasoning is required to obtain a correct answer here. But if the student gets the answer wrong, the system could begin by tutoring on the topic in the easier scenario as follows: "Let's think of how this relates to the previous question then. To begin with, in the case where the man exerted only a vertical force on the pumpkin, what was the

relationship between their horizontal displacements at all times after the release?" This way, we can verify that the student needs tutoring on a topic without implying that the student neglected to include the corresponding information in the student's essay. In the case where the student did include this information in the essay, the question can be interpreted as simply drawing the student's attention to something that was articulated earlier. Thus, this answer offers a potential solution both to the problem of neglecting to tutor on topics that are needed and to avoiding frustrating students by tutoring them on topics that they have already demonstrated competence at.

4 Conclusions and Current Directions

In this paper we have explored the issue of selecting appropriate interventions for students who start out with a high initial competency. We present a mathematical model that makes it possible to project success rate at selecting appropriate interventions based on accuracy at analyzing student performance at a task. This model demonstrates that for high initial competency students, selecting interventions based on any sort of isolated analysis of student performance is doomed to failure even with very high accuracy analysis. We explore an alternative approach to selecting interventions within the context of the WHY2 conceptual physics tutoring system (VanLehn et al., 2002) based on prior probabilities of student need. Our analysis demonstrates that within this domain, individual differences between high competency students makes this type of approach by itself equally unsuccessful. We are continuing to collect transcripts of students interacting with the WHY2 system and plan to explore the possibility of combining predictions based on prior probability of student need with predictions based on analysis as well as computing student specific prior probabilities based on performance on the pretest.

5 Acknowledgments

The authors would like to thank the rest of the Natural Language Tutoring group for their collaboration.

This research is supported by the Office of Naval Research, Cognitive and Neural Sciences Division MURI Grant N00014-00-1-0600 and NSF Grant 9720359 to CIRCLE, a center for research on intelligent tutoring.

References

K. VanLehn, P. Jordan, C. P. Rosé, and The Natural Language Tutoring Group. 2002. The architecture of why2-atlas: a coach for qualitative physics essay writing. In *Proceedings of the Intelligent Tutoring Conference.*