Computer Adaptive Testing Using the Same-Decision Probability

Suming Chen and **Arthur Choi** and **Adnan Darwiche**
Computer Science Department
University of California, Los Angeles
{suming,aychoi,darwiche}@cs.ucla.edu

Abstract

Computer Adaptive Tests dynamically allocate questions to students based on their previous responses. This involves several challenges, such as determining when the test should terminate, as well as which questions should be asked. In this paper, we introduce a Computer Adaptive Test that uses a Bayesian network as the underlying model. Additionally, we show how the notion of the Same-Decision Probability can be used as an information gathering criterion in this context — to determine which further questions are needed and if so, which further questions should be asked. We show empirically that utilizing the Same-Decision Probability is a viable and intuitive approach for determining question selection in Bayesian-based Computer Adaptive Tests, as its usage allows us to ask fewer questions while still maintaining the same level of precision and recall in terms of classifying competent students.

1 INTRODUCTION

Computer Adaptive Tests have recently become increasingly popular, as they have the ability to adapt to each individual unique user (Vomlel, 2004; Almond, DiBello, Moulder, & Zapata-Rivera, 2007; Almond, Mislevy, Steinberg, Yan, & Williamson, 2015). This in turn allows for the test to tailor itself specifically based on user responses and its current estimation of the user's knowledge level. For example, if the user answers a series of questions correctly, the test can adjust and curate some more difficult questions. On the other hand, if the user answers a series of questions incorrectly, the test can adjust and present easier questions.

Bayesian networks have been used as the principal base model for many Computer Adaptive Tests (Millán & Pérez-De-La-Cruz, 2002; Vomlel, 2004; Munie & Shoham, 2008; Almond et al., 2015), as they offer powerful approaches to inferring a user's knowledge level given uncertain information (unanswered questions). Additionally, a significant body of work has been done on using Bayesian networks to perform a wide array of tasks in the field of educational diagnosis (VanLehn & Niu, 2001; Conati, Gertner, & VanLehn, 2002; Suebnukarn & Haddawy, 2006).

A key question in this domain is when the test should terminate. Some students may perform so well or so poorly that the system can recognize that further testing is unnecessary, as asking further questions would have very little probability of reversing the initial diagnosis. (Millán & Pérez-De-La-Cruz, 2002) discusses some stopping criteria that are used to determine when further questioning is needed, as well as some selection criteria that are used to determine which questions are actually asked.

In this paper, we discuss the creation of a Computer Adaptive Test, and then take a recently introduced notion, called the *Same-Decision Probability* (SDP) (Darwiche & Choi, 2010), and show its usefulness as an information gathering criteria in this domain, in contrast to standard criteria. The SDP quantifies the stability of threshold-based decisions in Bayesian networks and is defined as the probability that a current decision would stay the same, had we observed further information.

The paper is structured as follows: We first present some motivation for the constructed Computer Adaptive Test. We then discuss some related work in the educational diagnosis field. Following that, we then show how the Same-Decision Probability can be used as a stopping and selection criterion for our Computer Adaptive Test. Finally, we discuss experiment setup, present empirical results demonstrating the usefulness of the Same-Decision Probability, and then conclude the paper.

2 MOTIVATION

The Berkeley Free Clinic[1] is a clinic staffed entirely by volunteers and offers medical and dental treatment for the sur-

[1] http://www.berkeleyfreeclinic.org/

rounding community. In particular, the dental section of the clinic offers a wide variety of services ranging from cleanings, x-ray services, fillings, and extractions. Due to the expertise required for these services, the volunteers need to be highly trained to apply their knowledge in a practical setting. The volunteers are thus required to undergo a training period in order to prepare them to serve at the clinic.

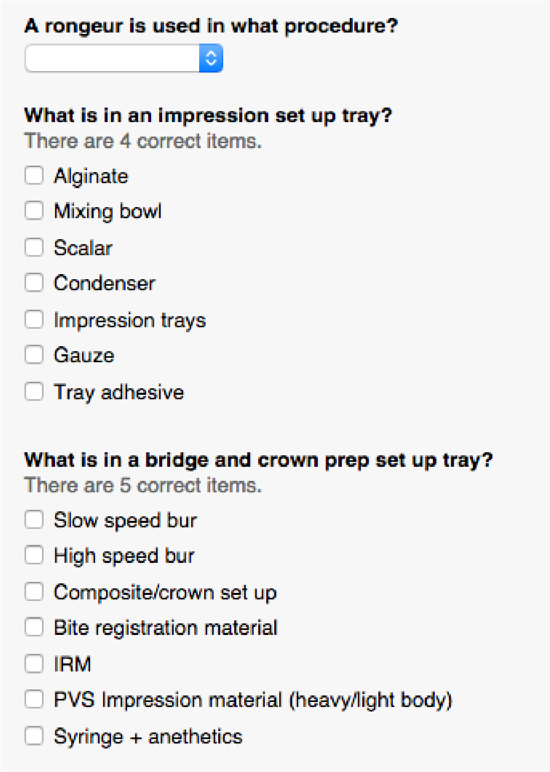
Recently, it was decided that in order to better evaluate the knowledge of the volunteers, the volunteers would go through a *competency exam* to determine whether or not they had sufficient knowledge. The idea of this test is that if a volunteer was found to be inadequate, that he/she would have to undergo further instruction. The coordinators of the clinic worked in conjunction with the volunteer dentists in order to create a written test that thoroughly tested the different concepts and skills necessary for a volunteer. The written test consists of 63 questions that includes questions ranging from showing a portrait of a certain instrument (e.g. a perioprobe) and asking "Name this instrument", to "What is in a bridge and crown set up tray?". A portion of this test can be seen in Figures 1, 2, 3, and 4.

This test was given to 22 subjects. Sixteen of the subjects were evaluated prior to the test-taking to be competent volunteers, whereas the other 6 subjects were evaluated to be non-competent volunteers. The clinic coordinators set the pass threshold to be at 60%, meaning that volunteers needed to correctly answer 60% of the questions in order to be considered competent. The test proved to be effective in that of the 16 competent volunteers, 15 of them passed, and of the 6 non-competent volunteers, none of them passed. The test taking duration ranged from 30 minutes to 1 hour. Feedback from the participants indicated that they felt this test was fair and covered the necessary bases to be a volunteer, but that the test was too time-consuming.

The complaints about the duration of the test is in line with the discoveries of (García, Amandi, Schiaffino, & Campo, 2007), who discover that a significant problem with web-based tests is that they are too long, which may hinder students' ability to perform as they simply become bored and careless. This serves as the chief motivation for our work, as we want to turn this test into a Computer Adaptive Test (CAT), so that we can present students a test with fewer questions, but just as many *relevant* questions — this way we can decrease overall test duration without compromising the effectiveness of the test.

3 RELATED WORK

There has been a surge of interest in applying artificial intelligence techniques in educational diagnosis. Educational diagnosis a broad field that covers Interactive Tutoring Systems (ITS) (Conati et al., 2002; Suebnukarn & Haddawy, 2006; Gujarathi & Sonawane, 2012), as well as Computer Adaptive Tests (CAT) (Munie & Shoham, 2008; Millán, Descalco, Castillo, Oliveira, & Diogo, 2013). When compared to traditional techniques, these applications have been shown to be highly effective in increasing the efficiency of student learning (Millán & Pérez-De-La-Cruz, 2002; Vomlel, 2004; Sinharay, 2006; Brusilovsky & Millán, 2007; Beal, Arroyo, Cohen, Woolf, & Beal, 2010; Millán et al., 2013).



A rongeur is used in what procedure?

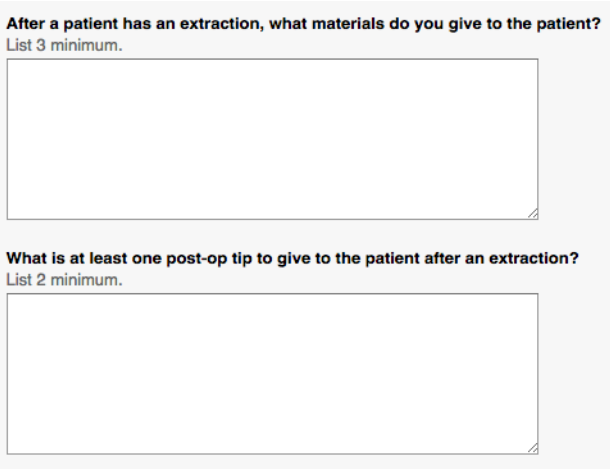
What is in an impression set up tray?
There are 4 correct items.
□ Alginate
□ Mixing bowl
□ Scalar
□ Condenser
□ Impression trays
□ Gauze
□ Tray adhesive

What is in a bridge and crown prep set up tray?
There are 5 correct items.
□ Slow speed bur
□ High speed bur
□ Composite/crown set up
□ Bite registration material
□ IRM
□ PVS Impression material (heavy/light body)
□ Syringe + anethetics

Figure 1: A portion of the competency test that measures general dental clinic knowledge.

In the work of (Millán & Pérez-De-La-Cruz, 2002), they use Bayesian networks as the framework for constructing Computer Adaptive Tests (CATs). They introduce a model for representing student knowledge, where knowledge is modeled as different interrelated concepts. They noted that student knowledge can be diagnosed (inferred) by treating student answers as evidence. In addition to this model, they introduced an adaptive testing algorithm. To evaluate their model, they used 180 simulated students. They found that the introduction of adaptive question selection improves both accuracy and efficiency. Their model incorporates notions such as "slipping", where a student might answer a question incorrectly even if the concept is known. They generate students randomly by sampling from the network, where each student is assumed to hold knowledge of different concepts.

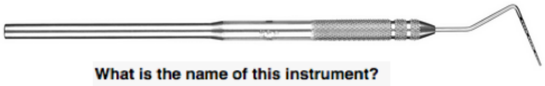
Similar results were shown in (Vomlel, 2004), who discuss applications of Bayesian networks in educational diagnosis, especially in skill diagnosis. They show that modeling



After a patient has an extraction, what materials do you give to the patient?
List 3 minimum.

What is at least one post-op tip to give to the patient after an extraction?
List 2 minimum.

Figure 2: A portion of the competency test that measures procedural knowledge.



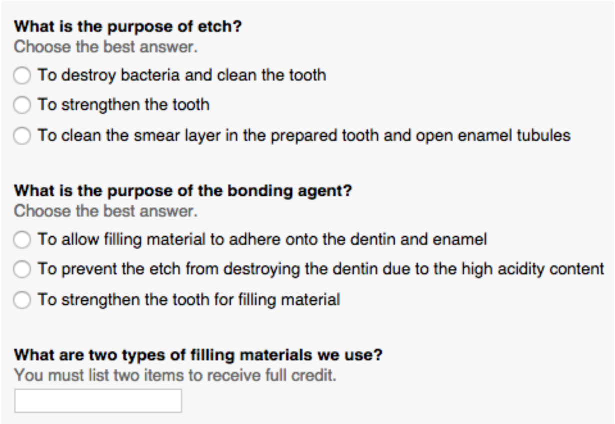
What is the name of this instrument?

Figure 3: A portion of the competency test that measures tool recognition knowledge.

dependence between various skills allows for higher quality of diagnosis. Additionally, they show that Computer Adaptive Tests allow them to best select a fitting question to test for competency in certain skills. They found that computer adaptive testing substantially reduces the number of questions that may need to be asked.

Just like (Millán & Pérez-De-La-Cruz, 2002), (Vomlel, 2004) and (Almond et al., 2015) found that Bayesian networks to be useful to model the links between various related proficiencies, and modeled unseen questions as unobserved variables. They use diagnostic Bayesian networks in order to model uncertainty, where experts build a Bayesian network and calibrate it from data.

Another application of Bayesian networks to model student knowledge is discussed in (Munie & Shoham, 2008). They use a Bayesian network to model qualifying exams for graduate students at Stanford, where their goal is to select the questions (observations) that can best measure the knowledge in order to determine whether or not the student should pass. They also use a threshold function and stress that reducing the number of questions asked is important. They also prove the NP-hardness of deciding an optimal set of questions to assign a student.



What is the purpose of etch?
Choose the best answer.

○ To destroy bacteria and clean the tooth
○ To strengthen the tooth
○ To clean the smear layer in the prepared tooth and open enamel tubules

What is the purpose of the bonding agent?
Choose the best answer.

○ To allow filling material to adhere onto the dentin and enamel
○ To prevent the etch from destroying the dentin due to the high acidity content
○ To strengthen the tooth for filling material

What are two types of filling materials we use?
You must list two items to receive full credit.

Figure 4: A portion of the competency test that measures conceptual knowledge.

Additionally, a variety of work has been done on using Bayesian networks in the educational diagnosis field for threshold-based decision making under uncertainty, such as 1) in (Xenos, 2004), where the authors model the educational experience of various students and use threshold-based decisions to determine whether or not a student is likely to fail, 2) in (Arroyo & Woolf, 2005), where the goal is to determine the type of learner a student was (in terms of attitude) 3) in (Gertner, Conati, & VanLehn, 1998), where the goal is to infer what part of a problem a student is having trouble with, in order to provide hint-based remediation, 4) in (Butz, Hua, & Maguire, 2004), where the constructed ITS can assess knowledge using a BN as well as recommending certain remediation, and a threshold-based decision is used to determine if a student is competent in an area or not.

Some other work in the field of educational diagnosis is found in (Brusilovsky & Millán, 2007), which describes a model that details the relationship between the knowledge of certain concepts and how the student will perform on certain questions. They note that if a user demonstrates lack of knowledge, the model can be used to locate the most likely concepts that will remedy the situation. They found that using a Bayesian model allows them to compute the probability of a student answering a question correctly given competency in some domain.

More recently, the work on educational diagnosis modeling has focused on even more intricate modeling and remediation methods. For instance, (Rajendran, Iyer, Murthy, Wilson, & Sheard, 2013) has focused on modeling user emotional states, in order to detect when users are frustrated as well as the cause of the frustration. By finding the root cause of the frustration, special measures may be taken to remediate. Similarly, (Chrysafiadi & Virvou, 2013) has done significant work in modeling a student's performance, progress, and behavior. The developed e-learning system has been deployed into production. (Chrysafiadi & Virvou, 2014) finds that modeling a user's detailed characteristics, such as knowledge, errors, and motivation can improve the quality of a student's learning process as it allows for better remediation. The need for personalized remediation is also further stressed as (Vandewaetere & Clarebout, 2014) studies the importance of giving learners instruction and support directly. They stress that learner models need to be adjusted and updated with new information about the learners knowledge, effective states, and behavior in order to be maximally effective.

4 COMPUTER ADAPTIVE TESTING

4.1 USING A BAYESIAN NETWORK

To address the problem of having such a time-consuming test, we believe that using a Computerized Adaptive Test (CAT) could differentiate the competent students from the non-competent students with *fewer* questions than a standard test. Asking fewer questions while still accurately measuring a student's ability is a canonical problem in educational diagnosis, and is discussed thoroughly in (Millán & Pérez-De-La-Cruz, 2002; García et al., 2007; Munie & Shoham, 2008; Millán et al., 2013). Our main goal is to best measure whether or not a student is competent with a *limited* subset of questions. This means that we have to determine when enough questions have been asked, as well as *which* additional questions should be asked, so as to ensure that overall the Computer Adaptive Test is comprised of fewer questions.

Bayesian networks have been found to be especially useful for decision making under uncertainty (Nielsen & Jensen, 2009). We believe that they provide us a natural mechanism to model both 1) student knowledge and 2) questions that may be potentially asked to measure student knowledge. In these models, Bayesian networks can model the relationships between various proficiencies. These models detail the relationship between knowledge of certain concepts and how the student will perform on certain questions.

Using a Bayesian network thus allows us to predict how likely a student is to have knowledge in a certain concept based on the student's answers. Hence, we worked with the clinic coordinators and experts to create a Bayesian network structure and elicit the parameters for this exam. Our developed model is similar to the models developed by (Millán & Pérez-De-La-Cruz, 2002; Vomlel, 2004; Almond et al., 2007; Munie & Shoham, 2008; Brusilovsky & Millán, 2007; Millán et al., 2013), where Bayesian networks are used to evaluate and in a sense, "diagnose" a student's degree of knowledge or competency.

In our network, we have a main variable of interest that is representative of the student's total level of knowledge. We refer to this variable as the *decision variable* (D), and it serves as an overall measure of our belief that a student is competent. General competency is determined by competency in a collection of specialized fields of knowledge, which are represented by a variety of latent *concept* variables. The final type of variable is a *question* variable that represents a question that may be asked to the student. An answered question, whether correct or incorrect, will act as *evidence* that can influence our belief on the student's competency. The graphical structure of the model can be seen in Figure 5.

Similarly to (Gertner et al., 1998; VanLehn & Niu, 2001; Cantarel, Weaver, McNeill, Zhang, Mackey, & Reese, 2014), the clinic coordinators/experts determined that the "pass threshold" should be set at 0.8, meaning that if given some evidence (questions answered), the *posterior* probability of the decision variable was found to be over 0.8, then the student would be considered as competent.[2] According to this pass threshold, we found once again that of the 16 competent volunteers, 15 of them passed, and of the 6 non-competent volunteers, none of them passed. The volunteers' scores are shown in Table 1. From this table we can see that our constructed Bayesian model allows us to accurately predict a student's competency.

Student #	Percentage Correct	Posterior Probability
1	0.952	0.957
2	0.921	0.941
3	0.714	0.951
4	0.539	0.704
5	0.762	0.956
6	0.746	0.950
7	0.825	0.957
8	0.619	0.948
9	0.777	0.952
10	0.857	0.954
11	0.857	0.957
12	0.809	0.956
13	0.349	0.153
14	0.238	0.023
15	0.539	0.178
16	0.809	0.923
17	0.603	0.883
18	0.635	0.954
19	0.873	0.954
20	0.524	0.135
21	0.492	0.153
22	0.413	0.314

Table 1: This table shows, for each student, the percentage of questions answered correctly (out of 63 questions), and the posterior probability of the student being competent given all the answered questions. The non-competent volunteers are bolded.

[2] Another commonly used threshold in the educational diagnosis field is 0.7 (Butz et al., 2004; Arroyo & Woolf, 2005).

4.2 STOPPING CRITERION FOR INFORMATION GATHERING

Although the test has a total of 63 questions, there may be a point during the adaptive test where we can determine that it is unnecessary to ask any further questions. For educational diagnosis in Bayesian networks, (Millán & Pérez-De-La-Cruz, 2002; Millán et al., 2013) shows there are two standard ways to determine if more information gathering (additional questions posed) is necessary. The first involves a fixed test that asks a set number of questions, and then determines if the student is competent after *all* questions have been answered. This static method is clearly ineffective, as the student is only evaluated at the *conclusion* of the test.

The second way involves terminating the test once the posterior probability of the decision variable is above or below some threshold. This stopping criteria is also seen in (Hamscher, Console, & de Kleer, 1992; Heckerman, Breese, & Rommelse, 1995; Kruegel, Mutz, Robertson, & Valeur, 2003; Lu & Przytula, 2006). Note that (Millán & Pérez-De-La-Cruz, 2002) compares an approach that utilizes a set number of questions with a threshold-based approach — they found that with a set number of questions they were able to diagnose students correctly 90.27% of the time, whereas by using an adaptive criterion they were able to diagnose students correctly 94.54% of the time, while requiring fewer number of questions to be asked. This is a clear indication that using a less trivial stopping criterion can be ultimately beneficial.

Another possibility of a stopping criterion involves computing the value of information of some observations, and if that value is not high enough, to then stop information gathering. However, this involves either computing the *myopic* value of information and just computing the usefulness of making one observation, or computing the *non-myopic* value of information and computing the usefulness of making several observations. The former is fairly easy to compute, but can prove to be very limited as often the combined usefulness of some observations is greater than its parts. For example, if a student answers a question incorrectly, it may not be very telling. However, if that student also answers a question incorrectly for an entirely orthogonal field, these two mistakes combined may be indicative of a student's lack of competency. Note that computing the *non-myopic* value of information is useful in determining if a set of observations has significant value, but is intractable to compute if the set of observations is large (Krause & Guestrin, 2009).

To decide whether or not enough information has been gathered, we use a non-myopic stopping criterion and compute the *Same-Decision Probability* (SDP) (Darwiche & Choi, 2010; Chen, Choi, & Darwiche, 2014).

Definition 1 *Suppose we are making a decision based on whether $Pr(D{=}d \mid \mathbf{e}) \geq T$ for some evidence $\mathbf{e}$ and threshold T. If $\mathbf{H}$ is a set of variables that are available to observe, then the SDP is:*

$$\mathsf{SDP}(d, \mathbf{H}, \mathbf{e}, T) = \sum_{\mathbf{h}} [Pr(d \mid \mathbf{h}, \mathbf{e}) \geq T] Pr(\mathbf{h} \mid \mathbf{e}). \quad (1)$$

Here, $[\alpha]$ is an indicator function which is 1 if α is true, and 0 otherwise.

In short, the SDP is a measure of how robust a decision is with respect to some unobserved variables and will help determine how much more information is necessary.

In this case, based on the student's responses, at any point in time, we can compute the posterior probability that they are competent and thus determine a decision of whether or not they are competent. Keep in mind that this decision is temporary and can be reversed if more questions are answered — we can compute the SDP over the remaining unanswered questions to determine how likely it is that the current decision (deciding whether a student is competent or non-competent) would remain the same even if the remaining questions were answered. If the SDP is high, that is an indication that we can terminate the test early. We found that using the SDP allowed us to cut the test duration significantly while maintaining diagnosis accuracy. Details of our experiment setup and results can be found in Section 5.1.

4.3 SELECTION CRITERION FOR INFORMATION GATHERING

We have now motivated the usefulness of the SDP as a stopping criterion for Computer Adaptive Tests, as it tells us when we can terminate the test. However, the question remains: if computing the SDP indicates that more questions are necessary, *which* questions should we ask?

In a standard, non-adaptive test, the ordering of the questions cannot be controlled. In the adaptive setting we have more control — based on a test taker's answers on the test so far, we can select questions that have the most potential to give us further insight on the test taker. Our goal here is to select questions such that the *expected* SDP after observing those questions is maximal. The expected SDP (Chen, Choi, & Darwiche, 2015) is defined as the following:

Definition 2 *Let $\mathbf{G}$ be a subset of the available features $\mathbf{H}$ and let $D(\mathbf{ge})$ be the decision made after observing features $\mathbf{G}$, i.e., $D(\mathbf{ge}) = d$ if $Pr(d \mid \mathbf{g}, \mathbf{e}) \geq T$. The E-SDP is then:*

$$\begin{aligned}&\mathsf{E\text{-}SDP}(D, \mathbf{G}, \mathbf{H}, \mathbf{e}, T)\\ &\quad = \sum_{\mathbf{g}} \mathsf{SDP}(D(\mathbf{ge}), \mathbf{H} \setminus \mathbf{G}, \mathbf{ge}, T) \cdot Pr(\mathbf{g} \mid \mathbf{e}). \quad (2)\end{aligned}$$

The expected SDP is thus a measure of how similar our decision, after observing $\mathbf{G}$, is to the decision made after observing $\mathbf{H}$. We want to find a question that will maximize the expected SDP. In other words, we want to ask questions such that no matter how they are answered, will, on average, minimize the usefulness of any *remaining* questions. By maximizing this objective, we reduce the overall number of questions that need to be answered before test termination.

As stated before, computing the *value of information* (VOI) is essential to the process of information gathering. Since for our model it is intractable to compute the non-myopic value of information, that leaves only computing the myopic VOI as an available possibility. The VOI of observing a variable may depend on various objective functions, for instance, how much the observation reduces the entropy of the decision variable. There is a comprehensive overview of these different objective functions in (Krause & Guestrin, 2009).[3] (Chen et al., 2014, 2015) compares the approach maximizing these standard objective functions to maximizing the expected SDP and finds that maximizing the expected SDP is more effective in reducing the number of questions that need to be answered, particularly in cases when the decision threshold is extreme.

Our approach involves selecting the question that leads to the highest expected SDP. We found that selecting variables to optimize the expected SDP allowed us to substantially decrease the number of questions selected compared to other selection criteria. Details of experiment setup and experimental results can be found in Section 5.2.

5 EXPERIMENTS

In this section, we empirically evaluate the SDP as a stopping and selection criterion for our adaptive test. We compare the SDP against the standard criteria used by other Bayesian Computer Adaptive Tests. First, we introduce some notation used throughout the experimental section.

We use standard notation for variables and their instantiations, where variables are denoted by upper case letters (e.g. X) and their instantiations by lower case letters (e.g. x). Sets of variables are then denoted by bold upper case letters (e.g. $\mathbf{X}$) and their instantiations by bold lower case letters (e.g. $\mathbf{x}$). The primary decision variable, measuring a student's overall competency, is denoted by D, with two states $+$ (competent) and $-$ (non-competent).

We have the completed tests for 22 subjects, where for each subject, we have the responses of the $n = 63$ test questions (to be more precise, we know whether each question was answered correctly or incorrectly). We denote this dataset by $\mathbf{T}$, where:

$$\mathbf{T} = \{\mathbf{e}_1, \ldots, \mathbf{e}_{22}\},$$

and where each $\mathbf{e}_i \in \mathbf{T}$ is an instantiation of the n test responses for student i. Hence, the probability

$$Pr(D = + \mid \mathbf{e}_i)$$

denotes the posterior probability that student i is competent given their test results. Using our Bayesian network, we decide that student i is competent if $Pr(D = + \mid \mathbf{e}_i) \geq 0.80$ (otherwise, we decide that they are not sufficiently competent). We report the quantity $Pr(D = + \mid \mathbf{e}_i)$ for each student, in Table 1.

To evaluate the SDP as a stopping and selection criterion, we simulate partially-completed tests from the fully-completed tests $\mathbf{T}$. In particular, we take the fully-completed test results $\mathbf{e}_i$, for each student i, and generate a set of partially-completed tests $\mathbf{Q}_i = \{\mathbf{q}_{i,1}, \ldots, \mathbf{q}_{i,n}\}$. In particular, we randomly permute the questions, and take the first j questions of the permutation as a partially-completed test $\mathbf{q}_{i,j}$. Hence, a partially-completed test $\mathbf{q}_{i,j+1}$ adds one additional test question to test $\mathbf{q}_{i,j}$, and test $\mathbf{q}_{i,n}$ corresponds to the fully-completed test $\mathbf{e}_i$.

In our experiments, we use those partially-completed tests $\mathbf{q}_{i,j}$ that have at least 10 questions, i.e., where $j \geq 10$ (we assume 10 questions to be the minimum number of questions, where we can begin to evaluate the competency of a student). Moreover, for each of the 22 students i, we simulated 50 sets of partially-completed tests $\mathbf{Q}_i$ based on 50 random permutations, giving us a total of $22 \cdot 50 = 1,100$ sets of partially-completed tests.

5.1 STOPPING CRITERION EXPERIMENTS

Using our partially-completed tests, we evaluate the SDP as a stopping criterion, against more traditional methods. In particular, we take each set of partially-completed tests $\mathbf{Q}_i$, and going from test $\mathbf{q}_{i,10}$ up to test $\mathbf{q}_{i,n-1}$, we check whether each stopping criterion is satisfied, i.e., a decision is made to stop asking questions. Note that given test $q_{i,n}$, the only decision is to stop, since there are no more test questions to ask. For the SDP, when we evaluate the test $\mathbf{q}_{i,j}$, we compute the SDP with respect to the remaining $n - j$ unanswered questions (i.e., we treat them as the set of available observables $\mathbf{H}$).

As for other, more traditional, stopping criteria, we consider (1) stopping after a fixed number of questions have been answered (after which a student's competence is determined), or (2) stopping early once the posterior probability $Pr(D = + \mid \mathbf{q}_{i,j})$ surpasses a given threshold T, after which we deem a student to be competent.

[3] Additionally, (Golovin & Krause, 2011) studies the usage of adaptive submodularity for selection criteria and shows that objective functions that satisfy this notion can be readily approximated. For our problem, our objective function does not satisfy the notion of adaptive submodularity.

In Table 2, we report the results where our stopping criterion is based on asking a fixed number of questions. In Table 3, we highlight the results where we use instead a posterior probability threshold (Millán & Pérez-De-La-Cruz, 2002; Xenos, 2004; Arroyo & Woolf, 2005; Munie & Shoham, 2008). These results are based on averages over our $1,100$ sets of partially-completed tests $\mathbf{Q}_i$, which we simulated from our original dataset. In both tables, we see that there is a clear trade-off between the accuracy of the test, and the number of questions that we ask. (Note again that students were already evaluated to be competent/non-competent prior to the test, and that precision/recall is based on this prior evaluation).

# Questions	Precision	Recall
10	0.566	0.403
15	0.752	0.659
20	0.805	0.755
25	0.815	0.812
30	0.835	0.835
35	0.874	0.875
40	0.901	0.909
45	0.915	0.914
50	0.918	0.926
55	0.924	0.938
60	0.950	0.938
62	0.954	0.938

Table 2: Precision and recall for a set number of questions.

T	Precision	Recall	# Questions Asked
0.750	0.803	0.758	10.13
0.775	0.813	0.821	13.04
0.800	0.836	0.832	16.02
0.825	0.867	0.872	19.26
0.850	0.903	0.915	23.13
0.875	0.919	0.912	27.14
0.900	0.908	0.911	31.57
0.925	0.968	0.884	36.82
0.950	0.974	0.818	44.25

Table 3: Precision, recall, and average number of questions asked, for varying thresholds T on the posterior probability.

We next consider the SDP as a stopping criterion. In particular, we compute the SDP with respect to all unanswered questions, and then make a stopping decision when the computed SDP surpasses a given threshold T. Note that when we use a threshold $T = 1.0$, then we commit to a stopping decision only when no further observations will change the decision (i.e., the probability of making the same decision is 1.0). In Table 4, we report the results of using the SDP as a stopping criterion.

We see that even when our threshold is set to a relatively small value (0.850), we still attain precision and recall rates that are comparable to those obtained by asking nearly all questions (as in Table 2). For the stopping criterion based on posterior probability thresholds (in Table 3), we can see that we can attain a higher precision, but at the expense of a lower recall. In fact, we see that once the threshold is set high enough, the recall actually drops: Some students who should be diagnosed as competent are in fact being diagnosed incorrectly as non-competent once the threshold is set too high.

Consider now the case where we set a very large threshold on the SDP (0.999). In this case, the precision and recall are equivalent to the case where our criterion is to ask a set number of $n-1 = 62$ questions (as in Table 2). In contrast, using the SDP criterion, we ask only an average of 42.05 questions, meaning that the SDP as a stopping criterion has the same robustness as the ”set number of questions” stopping criterion — while asking nearly 20 fewer questions.

SDP T	Precision	Recall	# Questions Asked
0.850	0.927	0.932	39.10
0.875	0.938	0.938	39.73
0.900	0.942	0.938	39.61
0.925	0.946	0.938	39.77
0.950	0.946	0.938	40.09
0.975	0.950	0.938	40.61
0.990	0.950	0.938	41.14
0.995	0.950	0.938	41.41
0.999	0.954	0.938	42.05

Table 4: Precision, recall, and average number of questions asked for different thresholds.

It is clear from our results that if our goal is to ask fewer questions, while maintaining competitive precision and recall rates, using the SDP as a stopping criterion is a compelling alternative to the traditional stopping criteria of 1) asking a set number of questions and 2) checking to see if the posterior probability of competency surpasses a threshold — using the SDP as a stopping criterion allows us to reduce the number of questions asked while still maintaining the same precision and recall.

5.2 SELECTION CRITERION EXPERIMENTS

We next consider experiments that compare various question selection criteria such as 1) random selection, where we select the next question randomly (as in non-adaptive or linear testing), 2) information gain (mutual information) (Millán & Pérez-De-La-Cruz, 2002; Vomlel, 2004), and 3) margins of confidence (Krause & Guestrin, 2009). Note that as this model does not fit the decision-theoretic setting (there are no assigned utilities), we do not consider utility maximization as a selection criterion. In addition to the above, we evaluate the SDP as a selection criterion.

Since our adaptive test involves selecting one question at a time, our goal is to select a variable $H \in \mathbf{H}$ (corresponding to a question not yet presented) that leads to the greatest gain in the SDP, i.e., the *SDP gain*. Our selection criteria is based on asking the question which yields the highest SDP gain, using information gain as a tie-breaker, when multiple questions have the same SDP gain. A tie-breaker is needed as the SDP gain of observing a single variable can be zero. In this case, observing a single variable is not enough to change our decision (multiple observations may be needed). In our experiments, we compute the SDP gain using a branch-and-bound algorithm, as in (Chen, Choi, & Darwiche, 2013). The advantages of this algorithm are (1) it can prune the corresponding search space, and (2) it can cache and reuse intermediate calculations, resulting in significant computational savings.

We compare how quickly these selection criteria allow us to stop information gathering, each using the SDP as a stopping criterion; see Table 5. Here, we took partially-completed tests (all tests where at least 30 questions were answered), and used each selection criterion to select additional questions to ask, until the SDP dictates that we stop asking questions. On average, we find that (1) the random selection criterion asks an additional **8.18** questions, (2) information gain asks an additional **3.67** questions, (3) margins of confidence ask an additional **3.57** questions, and (4) our approach based on the SDP gain requires the fewest additional questions: **2.77**.

Note that while the random question selection criterion clearly has the poorest performance, there is a relatively modest improvement based on our approach, using the SDP gain, compared to the more common approaches based on information gain (Millán & Pérez-De-La-Cruz, 2002; Vomlel, 2004), and margins of confidence (Krause & Guestrin, 2009). Nevertheless, the benefits of our approach, as a selection criteria, are still evident. Further, the results suggest the potential of using the SDP gain in a less greedy way, where we ask *multiple questions* at a time, which we consider a promising direction for future work.

SDP T	Random	IG	Margins	SDP
0.850	4.77	2.98	2.71	2.05
0.875	4.78	3.08	2.87	2.14
0.900	5.08	3.20	3.12	2.20
0.925	5.95	3.19	3.04	2.31
0.950	6.81	3.37	3.44	2.47
0.975	7.86	3.73	3.56	2.96
0.990	10.64	4.19	4.29	3.18
0.995	12.95	4.32	4.22	3.43
0.999	14.80	4.97	4.91	4.25

Table 5: Number of additional observations necessary before stopping for the random selection criterion, information gain criterion, margins of confidence criterion, and the SDP hybrid criterion.

5.3 RUNNING TIMES

The SDP has been shown to be highly intractable, being $\mathrm{PP}^{\mathrm{PP}}$-complete (Choi, Xue, & Darwiche, 2012). Therefore, the computational requirements of the SDP, as a stopping and selection criterion, may be higher than other, more common approaches. This complexity depends largely on the number of unasked questions that must be considered. Figure 6 shows a plot of the running times for the SDP stopping and selection criteria, in comparison to other measures. Note that the SDP selection criterion is on average more efficient than the SDP stopping criterion, even though the selection criterion includes an SDP gain computation, as well as an information gain computation (as a tie breaker). Here, there are cases that can be detected that allow us to skip the SDP gain computation (Chen et al., 2013), leaving just the relatively efficient information gain computation. In general, we note that there is a trade-off: the SDP, as stopping and selection critera, provides valuable information (leading to fewer questions asked), but the SDP is also more computationally demanding.

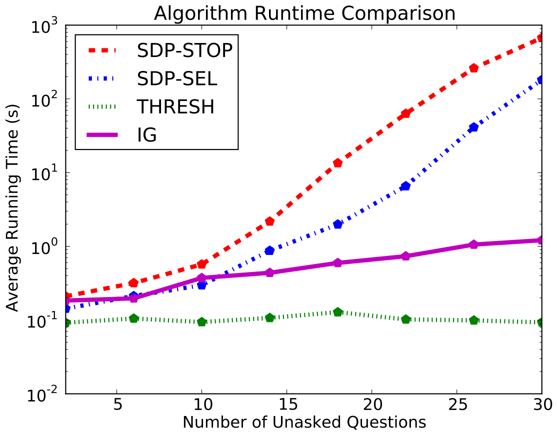


Figure 6: Runtimes of different algorithms. `SDP-STOP` and `SDP-SEL` are respectively the stopping and selection criterion algorithms, `THRESH` represents the standard posterior stopping criterion, and `IG` represents the standard information gain selection criterion.

6 CONCLUSION

We created a Computer Adaptive Test using a Bayesian network as the underlying model and showed how the notion of the SDP can be used as an information gathering criterion in this context. We showed that it can act as a stopping criterion for determining if further questions are needed, and as a selection criterion for determining which questions should be asked. Finally, we have shown empirically that the SDP is a valuable information gathering tool, as its usage allows us to ask fewer questions while still maintaining the same level of precision and recall for diagnosis.

Acknowledgments

This work has been partially supported by ONR grant #N00014-12-1-0423, NSF grants #IIS-1514253 and #IIS-1118122, and a Google Research Award. We also thank the coordinators and volunteers at the Berkeley Free Clinic.

References

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, *44*(4), 341–359.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian Networks in Educational Assessment*. Springer.

Arroyo, I., & Woolf, B. (2005). Inferring learning and attitudes from a Bayesian network of log file data. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, pp. 33–40.

Beal, C. R., Arroyo, I., Cohen, P. R., Woolf, B. P., & Beal, C. R. (2010). Evaluation of Animalwatch: An Intelligent Tutoring System for arithmetic and fractions. *Journal of Interactive Online Learning*, *9*(1), 64–77.

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web*, pp. 3–53. Springer-Verlag.

Butz, C. J., Hua, S., & Maguire, R. B. (2004). A web-based Intelligent Tutoring System for computer programming. In *Web Intelligence*, pp. 159–165.

Cantarel, B. L., Weaver, D., McNeill, N., Zhang, J., Mackey, A. J., & Reese, J. (2014). Baysic: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC bioinformatics*, *15*(1), 104.

Chen, S., Choi, A., & Darwiche, A. (2013). An exact algorithm for computing the Same-Decision Probability. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pp. 2525–2531.

Chen, S., Choi, A., & Darwiche, A. (2014). Algorithms and applications for the Same-Decision Probability. *Journal of Artificial Intelligence Research (JAIR)*, *49*, 601–633.

Chen, S., Choi, A., & Darwiche, A. (2015). Value of information based on Decision Robustness. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI)*.

Choi, A., Xue, Y., & Darwiche, A. (2012). Same-Decision Probability: A confidence measure for threshold-based decisions. *International Journal of Approximate Reasoning (IJAR)*, *2*, 1415–1428.

Chrysafiadi, K., & Virvou, M. (2013). Persiva: An empirical evaluation method of a student model of an intelligent e-learning environment for computer programming. *Computers & Education*, *68*, 322–333.

Chrysafiadi, K., & Virvou, M. (2014). Kem cs: A set of student's characteristics for modeling in adaptive programming tutoring systems. In *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on*, pp. 106–110. IEEE.

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, *12*(4), 371–417.

Darwiche, A., & Choi, A. (2010). Same-Decision Probability: A confidence measure for threshold-based decisions under noisy sensors. In *5th PGM*, pp. 113–120.

García, P., Amandi, A., Schiaffino, S., & Campo, M. (2007). Evaluating Bayesian networks precision for detecting students learning styles. *Computers & Education*, *49*(3), 794–808.

Gertner, A. S., Conati, C., & VanLehn, K. (1998). Procedural help in Andes: Generating hints using a Bayesian network student model. In *Proceedings of the 13th National Conference on Artificial intelligence*, pp. 106–111.

Golovin, D., & Krause, A. (2011). Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *JAIR*, *42*, 427–486.

Gujarathi, M. V., & Sonawane, M. S. (2012). Intelligent Tutoring System: A case study of mobile mentoring for diabetes. *IJAIS*, *3*(8), 41–43.

Hamscher, W., Console, L., & de Kleer, J. (Eds.). (1992). *Readings in Model-Based Diagnosis*. Morgan Kaufmann Publishers Inc.

Heckerman, D., Breese, J. S., & Rommelse, K. (1995). Decision-theoretic troubleshooting. *Communications of the ACM*, *38*(3), 49–57.

Krause, A., & Guestrin, C. (2009). Optimal value of information in graphical models. *Journal of Artificial Intelligence Research (JAIR)*, *35*, 557–591.

Kruegel, C., Mutz, D., Robertson, W., & Valeur, F. (2003). Bayesian event classification for intrusion detection. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*.

Lu, T.-C., & Przytula, K. W. (2006). Focusing strategies for multiple fault diagnosis. In *Proceedings of the 19th International FLAIRS Conference*, pp. 842–847.

Millán, E., Descalco, L., Castillo, G., Oliveira, P., & Diogo, S. (2013). Using Bayesian networks to improve knowledge assessment. *Computers & Education*, *60*(1).

Millán, E., & Pérez-De-La-Cruz, J. L. (2002). A Bayesian diagnostic algorithm for student modeling and its evaluation. *User Modeling and User-Adapted Interaction*, *12*(2-3), 281–330.

Munie, M., & Shoham, Y. (2008). Optimal testing of structured knowledge. In *Proceedings of the 23rd National Conference on Artificial intelligence*, pp. 1069–1074.

Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Rajendran, R., Iyer, S., Murthy, S., Wilson, C., & Sheard, J. (2013). A theory-driven approach to predict frustration in an its. *Learning Technologies, IEEE Transactions on*, *6*(4), 378–388.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, *31*(1), 1–33.

Suebnukarn, S., & Haddawy, P. (2006). A Bayesian approach to generating tutorial hints in a collaborative medical problem-based learning system. *Artificial intelligence in Medicine*, *38*(1), 5–24.

Vandewaetere, M., & Clarebout, G. (2014). Advanced technologies for personalized learning, instruction, and performance. In *Handbook of Research on Educational Communications and Technology*, pp. 425–437. Springer.

VanLehn, K., & Niu, Z. (2001). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. *International Journal of Artificial Intelligence in Education*, *12*(2), 154–184.

Vomlel, J. (2004). Bayesian networks in educational testing. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *12*(supp01), 83–100.

Xenos, M. (2004). Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education*, *43*(4), 345–359.

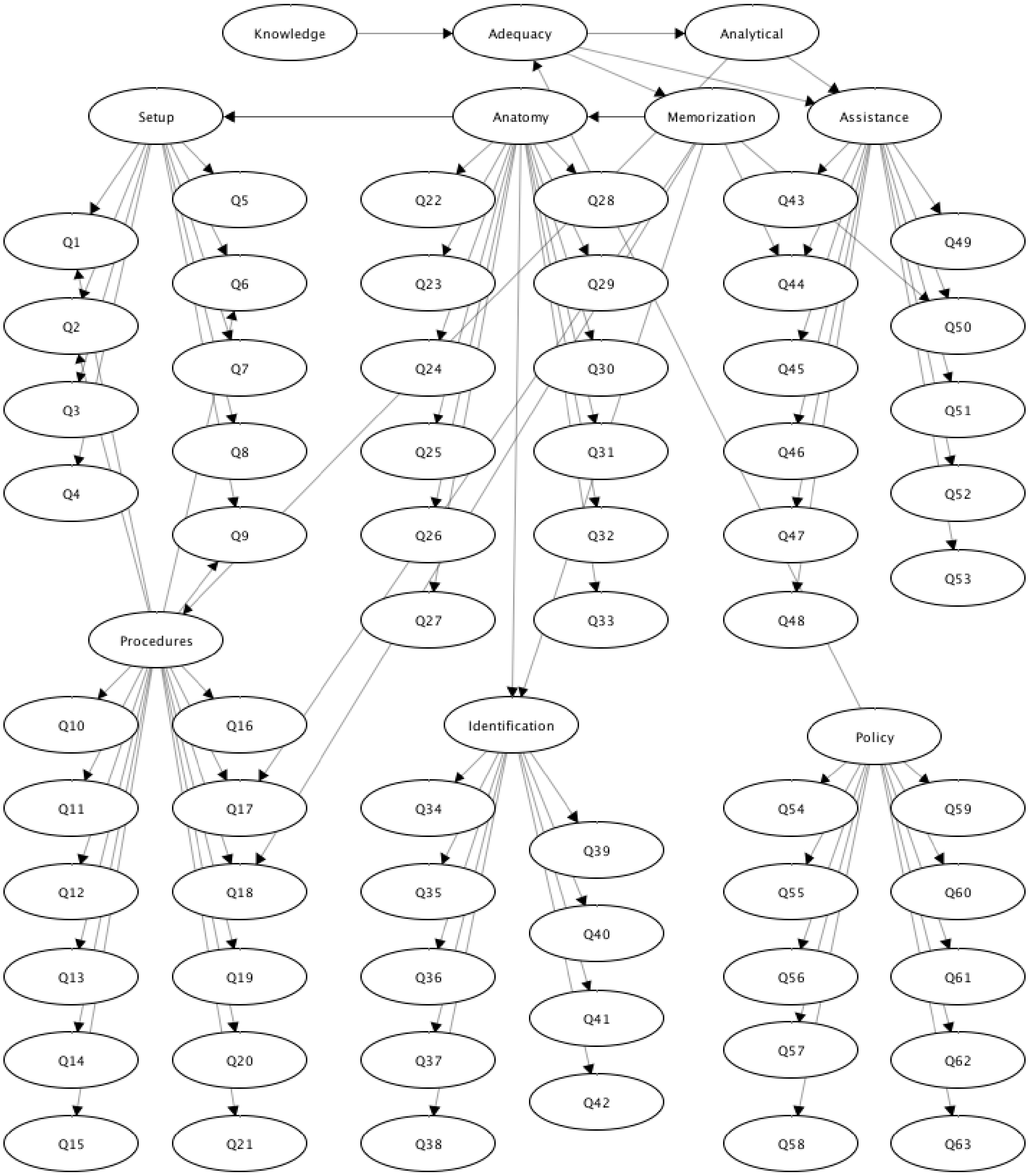


Figure 5: Bayesian network modeling clinic volunteer knowledge. The 63 questions are labeled from Q1 to Q63. The primary decision variable *Adequacy* is located in the top center.