Personalization Through the Application of Inverse Bayes to Student Modeling

Personalization through the Application of Inverse Bayes to Student Modeling

Charles William McLeod Lang

Prof. Howard Gardner
Prof. Terrence Tivnan
Prof. Ryan Baker

A Thesis Presented to the Faculty of the
Graduate School of Education of Harvard
University in Partial Fulfillment of the
Requirements for the Degree of Doctor of
Education

2015

Acknowledgements

I would like to thank my doctoral advisor, Howard Gardner, for his tireless support and guidance. I would also like to thank my dissertation committee, Terrence Tivnan, Ryan Baker and Katherine Masyn, without whose direction and knowledge this dissertation would not have been possible. I owe my deepest gratitude to Neil Heffernan and Yutao Wang whose generosity with both time and data was invaluable. I would also like to thank my parents Wendy and Douglas Lang, and the many colleagues whose help made this dissertation possible: Mariam Chughtai, Lissa Young, Lauren Elmore, Aaliyah El-Amin, Tiffanie Ting, Karin Liiv, Courtney Pollack and Raygine DiAquoi. A special thanks also goes to the good people of Stack Overflow, a marvel of the modern world.

Contents

Abstract

Personalization, the idea that teaching can be tailored to each students' needs, has been a goal for the educational enterprise for at least 2,500 years (Regian, Shute, & Shute, 2013, p.2). Recently personalization has picked up speed with the advent of mobile computing, the Internet and increases in computer processing power. These changes have begun to generate more and more information about individual students that could theoretically be used to power personalized education. The following dissertation discusses a novel algorithm for processing this data to generate estimates of individual level attributes, the Inverse Bayes Filter (IBFi).

A brief introduction to the use of Bayes Theorem is followed by a theoretical chapter and then two empirical chapters that describe alternately how the model is constructed, and how it performs on real student data. The theoretical chapter presents both the theory behind Inverse Bayes, including subjective probability, and then relates this theory to student performance. The first empirical chapter describes the prediction accuracy of IBFi on two proxies for students' subjective probability, partial credit and cumulative average. This prediction performance is compared to the prediction accuracy of a modified Bayesian Knowledge Tracing model (KTPC) with IBFi performing reasonably, out-predicting the KTPC model on a per-student basis but not across all predictions.

In the second empirical chapter I validate the theory behind the Inverse Bayes Filter through testing whether student certainty (or confidence) improves prediction performance. The inclusion of student certainty is shown to improve the predictive performance of the model relative to models that do not use certainty. This evidence supports the IBFi model and its underlying theory, indicating that students' judgments about their levels of certainty in their answers contains information that can be successfully identified by the model. A final summary chapter describes the consequences of using this model for education broadly.

Keywords: subjective probability, Bayes Theorem, Inverse Bayes Formula, Intelligent Tutors, assessment

Personalization through the Application of Inverse Bayes to Student Modeling: Introductory Bookend

Introduction

Since the eighteenth century Bayes' Theorem has gone from parlor game to powering the technological world. Originally devised by Thomas Bayes to calculate the proportion of colored marbles in urns, it now enables predictive text in mobile devices (Maragoudakis, Tselios, Fakotakis, & Avouris, 2002), missile defense systems (Tan, Wang, Shen, & Xu, 2005), quality control in factories (Singpurwalla, 1992) and the GRE adaptive test (Swinton, 1987). The following articles investigate whether the application of a particular flavor of Bayesian Analysis can help us automate the personalization systems for students. As background, the following bookend outlines the basic premise of the research - that Bayes Theorem can model the way that students make decisions and that this can provide insight into a) their learning and b) the impact of the context that this learning takes place in.

Bayes Theorem

Bayes Theorem describes the relationship between the probabilities of A and B, P(A) and P(B), and the conditional probabilities of A given B and B given A, P(A|B) and P(B|A):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

As a concrete example, you see a friend talking to someone across the street in Cambridge, MA. You think your friend may be speaking to a professor, but you are unsure. You observe that the person is wearing a tweed jacket though. In this instance you are attempting to calculate the conditional probability that your friend is speaking with a professor, given that the person in question is wearing a tweed jacket, P(Prof|Tweed Jacket). We can make this calculation provided we know:

- The probability that someone is a professor or our *prior knowledge* of seeing professors P(Professor)
- The probability of wearing tweed, P(Tweed Jacket) and
- The *likelihood* that someone wears tweed jackets given that they are professor, P(Tweed Jacket|Prof)

According to the Cambridge Community Development Department, 5% of those that live and work in Cambridge are professors, so P(Prof) is 0.05 (Cambridge Community Development Department, 2010). Suppose we also know something about the fashion habits of Cambridge generally. Tweed is a robust market in Cambridge with 1 in 10 people wearing a jacket regularly, and professors in particular being fond of the style with 2 in 5 wearing one regularly. We can then fill out the remaining probabilities in the calculation, P(Tweed Jacket) is equal to 0.10 and the probability of wearing a tweed jacket, given that you are a professor, P(Tweed Jacket|Professor), is 0.66. We can then estimate the posterior probability that the friend is having a discussion with a professor as:

$$p(professor|tweed\ jacket) = \frac{0.66 \times 0.05}{0.10} = 0.33 \qquad (2)$$

Not a very high probability at all, this is the *posterior probability*, as it conceptually occurs after the likelihood and prior have been multiplied together. It is also referred to variously as the *logical probability* or *reasonable estimation*. Neither of these terms should invoke the sense that it is the best value though, as it is entirely dependent on the values of the prior probability and likelihood – a poor choice of these values will produce a misleading result. Far from being a weakness of the Theorem though, this property of Bayes allows us to compare different estimates of these values. The comparison of values is the first idea that we will attempt to employ to make inferences about students: strength of belief. For example, aside from situations involving tweed coats and Cambridge professors, we can apply Bayes' theorem to situations involving students' problem solving. In looking at a question on a test, for example, there is a probability that a student will get the item correct (if it is a difficult item, the probability may be relatively low). Yet the student might show a high level of confidence in the answer, so this might change our thinking about the probability that the student gets the item correct. Using Bayesian reasoning can be a big help here, at least if we make reasonable assumptions about the prior probabilities and the likelihood

Model Modification

Flexible Estimates

The dominant understanding of probability is as a proportion, otherwise known as the Frequentist interpretation of probability. Within this interpretation it is possible to make an estimation of a *true* value of a given proportion through statistics. Alternately, in Bayesian inference, probability is considered a "strength of belief" and although it can be a proportion, it can also be considered a tendency or propensity (Popper, 1959) not reliant on having a long range frequency to generate a proportion from. Doing away with the necessity of propensity opens the door to modeling a much wider array of phenomena for which proportions are unavailable (e.g., situations in which it is unethical to collect data) or do not make sense (e.g., the proportion of all possible events in the universes).

The idea of strength of belief also allows estimates to be flexible and change over time as more evidence becomes available or context changes. Unlike frequency based statistical models of student behavior such as True Scores or IRT models that attempt to make estimates of the true value of students' knowledge or intelligence, a Bayesian estimate of student knowledge can change over time (Borsboom, 2005, p. 79). Change over time allows us to take advantage of two characteristics in investigating student behavior – comparison of prior probabilities and updating of likelihoods.

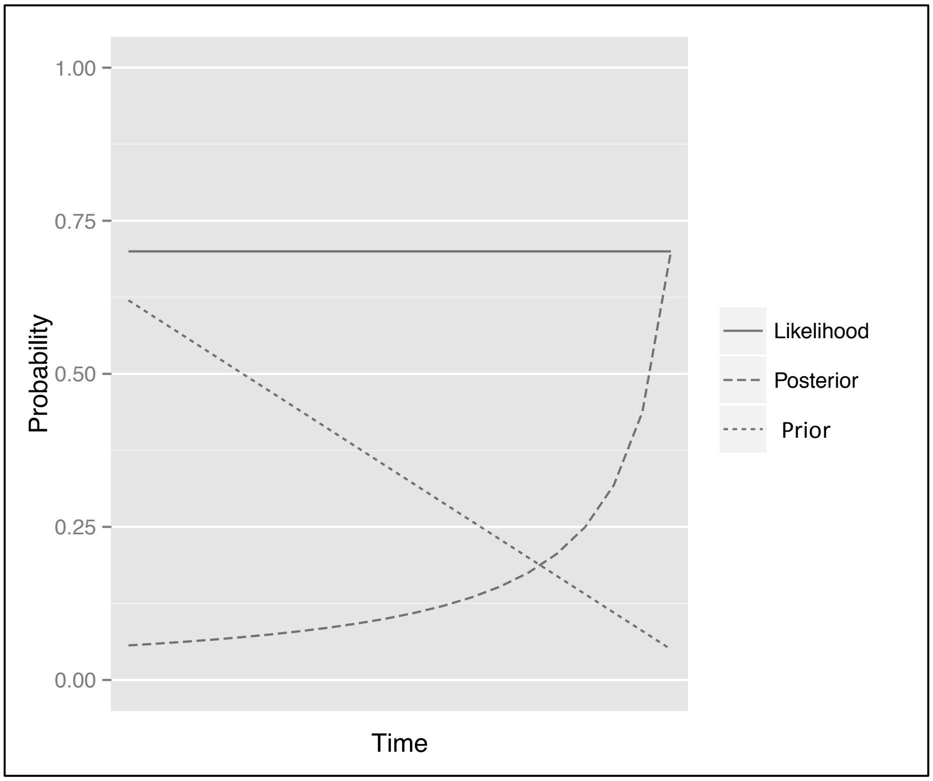


Figure 1. Changes in the popularity of tweed among the general population (P(Tweed)) are not reflected by the popularity of tweed among professors resulting in climbing predictions of whether or not someone is a professor if they are wearing tweed.

Bayesian Recursion. The basic model of likelihood and prior updating is Bayesian Recursion. For example, to improve our prediction of the professor, we might also factor in other, time dependent information such as changes in fashion. If we closely follow fashion trends we may see a fall in the popularity of tweed and we can model this by sequentially changing the P(Tweed). We can also model whether professors are resistant or susceptible to changes in fashion by altering the likelihood, P(Tweed|Professor). For example, perhaps we observe that among members of the general public tweed becomes less popular but professors continue to wear tweed despite this change in fashion. Given these factors we can see that over time our prediction of whether our friend is speaking to a professor increases (Figure 1).

Updating is especially useful when we are dealing with noisy data over time because we can continually refine our search for a signal. Consider there are many traits that professors tend to have, but none of them definitively tells us that someone is a professor: wearing tweed, carrying books, speaking to students, living close to a university, etc. If we remain interested in positively identifying our friend's companion as a professor we could follow her over time, noting how many of these traits she exhibits and updating our prediction accordingly (Figure 2).

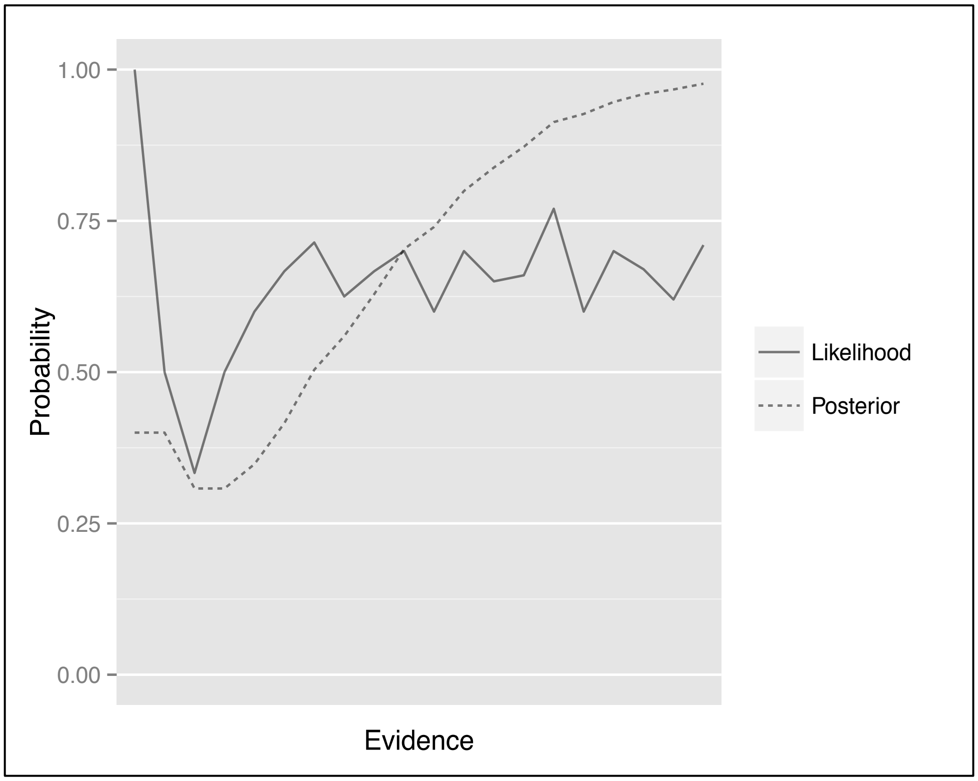


Figure 2. Changes in the posterior probability (i.e., – prediction) of whether you have identified a professor as you gather more evidence. Even though the proportion of evidence changes erratically, the prediction steadily grows as more confirmatory evidence is collected.

One system for making these updates is Bayesian Recursion. In Bayesian Recursion the likelihood is sequentially updated as new data become available and then the previous prediction becomes the new prior probability allowing us to generate a new posterior. Therefore, applying this method to our prediction problem, we could track the individual in question, noting professorial evidence over time until we reached a stable estimate of the probability.

Comparison of prior probabilities. To continue with our example of identifying a professor, let's say you see your friend and their companion walk onto a university campus where your prior probability of seeing a professor is much more likely, P(Professor) = 0.20. Such a change increases your prediction substantially:

$$p(professor|tweed\ jacket) = \frac{0.40 \times 0.20}{0.10} = 0.80 \tag{3}$$

In the above example we altered the context and so changed the prior information that we were using to calculate the probability of a professor. The ability to alter the prior means that alternate viewpoints or models of phenomena can be compared. For example, if we treat prior information to be a sense of expertise, we can compare the predictions of different experts. Modeling of this type has found use in climate science (Choy, O'Leary, & Mengersen, 2009), geology (Baddeley, Curtis, & Wood, 2004) and economics (Lombardi & Nicoletti, 2012) and it represents a shift from using Bayes to make the best, objective prediction, to using Bayes to represent the thoughts of people.

Bayes as Model of the Mind

Approaches

The conceptual change from using Bayes to model uncertainty to modeling prior knowledge has lead to a raft of attempts to use Bayes as a model of the human mind. Three such modes are Decision Theory, Bayesian Knowledge Tracing and Cognitive Bayesian approaches.

Decision theory. Decision Theory grew out of financial risk assessment and so is concerned with describing normative decision spaces in which there is a clear optimal outcome or *expected utility* (Ghirardato, 2002). Bayes has found a natural home in this field as a stand in for the optimal decision maker. Indeed, the Complete Class Theorems imply that all admissible decision rules are approximately Bayesian (Le Cam, 1955; Sacks, 1963; Stein, 1955; Wald, 1950).

Although successful theoretically, the optimum decision maker model does not fit actual human behavior well. Allais (1953) and Ellsberg (1961) and then Kahneman and Tversky (1972) showed reliably that humans tend to diverge from optimum Bayesian decision-making in predictable ways. Although this was not the final word, and Bayes did make some inroads within psychology there was not much research utilizing Bayesian methods in psychology until the end of the 1990s. In the meantime, Bayesian decision-making had been taken up with alacrity by both the machine learning community as a means of feature reduction and designers of Intelligent Tutors as a means for prediction.

Bayesian Knowledge Tracing. Bayesian Knowledge Tracing (BKT) is the dominant form of prediction used in Intelligent Tutoring Systems. First proposed in the late 1960s by Atkinson (1972) it was further developed by Corbett and Anderson (1994). BKT uses Bayes Theorem to derive estimates of the probability that a discrete skill has been learned, P(Ln). It is used to trace skill development in students as they work through problems in Intelligent Tutors. BKT was the most direct use of Bayes in an educational context until recently with the development of large, complex online learning environments that adopted machine-learning techniques.

Machine learning approaches: cognitive Bayes. Creating computers that could make the best decision within constrained hypothesis spaces was required to automate many industries. From car-building robots to Internet commerce, the demand for predictive technology and technology that could simplify complex phenomena made Bayes a useful tool.

As a result, partnerships in the early 2000s between machine-learning researchers and psychologists saw a small revival in using Bayes to model human behavior. Spearheaded by Gopnik (Gopnik, 2008; Gopnik et al., 2004; Gopnik & Glymour, 2002; Gopnik & Tenenbaum, 2007), this work models the causal reasoning of children by using Bayes nets to describe how people link events together. She showed that children, although they may not make decisions in keeping with Bayes, do describe causal relationships in accordance with Bayesian Principles. Griffiths and Tenenbaum extended this approach beyond causal reasoning to include induction (Tenenbaum, 2000), vocabulary learning (Frank, Goodman, & Tenenbaum, 2009), judging similarity (Tenenbaum, 1996), and forming perceptual representations (Griffiths & Warren, 2004). These "probabilistic models of cognition" seek to describe the *algorithmic* characteristics of human behavior rather than *implementational* (biological) or *computational* (experimental psychological) characteristics (Marr & Poggio, 1979). They are trying to infer that people can algorithmically rely on Bayes to generate inferences under certain conditions. The ability to invoke Bayes does not necessarily imply that people are rational or Bayesian by nature, but rather that Bayes is one tool that people have access to from an early age, in the same vein that developmental psychologists discuss the intuitive number sense (Feigenson, Dehaene, & Spelke, 2004).

Modifying Bayesian Models of the Mind

Modeling Traits

Although the idea that humans are fundamentally Bayesian has been continually questioned, Bayesian psychological models are still geared at uncovering general, human-level traits. Often they take the form of testing whether people do, or do not, on average, behave in a Bayesian fashion under a certain set of circumstances. The applicability of this approach to education research may be limited though. Although generalizability is a laudable aim, there may not be anything to generalize as student heterogeneity, particularly with respect to the way people understand new information, may be too great. In Bayesian probabilistic models of cognition, the assumption is that this is not the case. Rather, it is assumed that everyone processes incoming information in the same way and it is their prior knowledge that differentiates their understanding. For example, referring back to the baby and the sunrise, the assumption is that if there were several babies, they would all see the sunrise in the same way and therefore adopt new data with the same efficiency. Such a strategy may be reasonable in some circumstances, but in a classroom it is highly likely that students will receive new data with varying degrees of efficiency as well as having different prior knowledge. A student may not be looking in the direction of the teacher, she may be distracted, or she may have poor hearing – all these possibilities are clearly non-random and impact her ability to alter her beliefs.

As another example, a standard experiment might involve flipping a coin and asking people whether or not they think it is a fair coin that produces 50:50, heads:tails, or a trick coin that produces more heads than tails. The people in the trial may or may not, on average, update their decision about the coin in accordance with Bayes Theorem as they are shown more flips of the coin. However, the assumption of the model is that with each flip, each person will incorporate the new information into her or his schema with equal, if not perfect, accuracy. This experiment does not account for lapsed concentration or the differential sight of participants, for example; all those differences are considered random noise. As such, people will either behave in a Bayesian or non-Bayesian fashion. This strategy does not allow for the situation that people may behave in a Bayesian fashion, contingent upon how they interpret new information. This assumption is problematic for behavioral research generally, but in particular it is problematic for educational research. It may be reasonable to treat these factors as random in a laboratory setting, but in a classroom it a stretch to assume that distraction, student mood, or the temperature are random events.

The Bayesian Classroom. The bread and butter of the educator is not the human level trait - characteristics common to all humans. Of course, there are instances in which differences in fundamental memory or executive function are important, but the majority of lessons outside the special education classroom do not revolve around these fundamental differences. Rather, educators are trying to work with the differential mental representations, ideas, and understandings of their students (Siegler, 1996). Sometimes they are trying to homogenize ideas across students - everyone in the class should understand that 1 + 1 = 2, and sometimes they are trying to differentiate ideas – everybody should have their own understanding of the ethics of organ donation. Regardless, it is the differences between students that are being manipulated rather than fundamental human characteristics and a key source of this difference is how prior knowledge may impact the processing of new information.

The translation from machine learning to psychology, and any future adaption of Bayes for education, needs to take into account the heterogeneity of student data processing in a substantive way. When creating a robot that can kick a ball, it isn't necessary to understand why the machine does or does not perform the task, provided that the algorithm can be tweaked until a desirable result is achieved. The explanation for why it works in a psychological sense is unnecessary and the information about parameters is not likely to be human-readable anyway. Conversely, a teacher is required to understand why students behave the way they do. There are myriad reasons why teachers would want to describe what is going on with their students: to create student profiles, to justify interventions, or to promote self-efficacy among students.

Bayesian models do go part way toward describing the inner workings of people; they allow for the differential modeling of prior knowledge. They do not allow for the differential modeling of how students grasp new information, though. To alter Bayesian models to allow this kind of flexibility requires an adaption of Bayesian methods named Inverse Bayes.

Inverse Bayes

As described above, the standard use of Bayes involves the prediction of the posterior probability using a *reasonable* prior and a likelihood generated from data: e.g., the prediction of whether or not someone is a professor given that they are wearing tweed. Inverse Bayes is simply the inversion of this idea, the use of posterior probabilities to generate priors and likelihoods: e.g., ascertaining the proportion of people who are professors, based on predictions of whether a set of people are professors wearing, or not wearing, tweed. For example, I know the probability of being a professor, given you are wearing tweed, to be 0.70. If I also know the likelihood of wearing tweed, given you are a professor is 0.5 and the probability of wearing tweed is 0.1, then I can calculate the prior probability of being a professor:

$$p(professor|tweed\ jacket) = 0.70 = \frac{0.50 \times p(professor)}{0.10}$$

Things become slightly trickier if I am missing the likelihood information, but I will at least know the range of the likelihood because the posterior limits these values. I can then use other heuristic methods (covered in subsequent articles) to calculate the range of the prior.

Inversion of Bayes is not a common method, however, and general proofs for its use were only developed in the late 1990s (Tian & Tan, 2003). Its use is also geographically/culturally located mostly to Asia and Russia and is very rarely adopted in West. Tian and Tan attribute this low adoption to the philosophical defense of Bayesianism against Frequentist critique (p.306). In these discussions the prior probability came to be considered a sacred cow of sorts and the idea that it could be generated simply by running everything backwards seemed to undermine its status. Regardless, there is no mathematical or logical reason why Bayes Theorem cannot be inverted, although it has not been thoroughly empirically tested (Tian, Ng, & Tan, 2010).

Utilizing Inverse Bayes for educational applications may be particularly useful as it allows the modeling of differential student data acquisition. By running Bayes backwards, the likelihood no longer describes a standard presentation of data (e.g., everyone is considered to be equally adept at internalizing that a coin has been flipped heads), rather the value represents how well data are internalized. This characterization of the likelihood allows the characterization of the difference between students in terms of how they are responding to stimuli. Whereas the prior represents the knowledge acquired with previous experience, the likelihood represents the probability of the data given the context and way it is presented. Are the data in opposition or agreement with the student's prior knowledge? In this way the pairing of priors and likelihoods might be a reasonable proxy for learning.

The following three articles investigate whether it is fruitful for educational research to pursue this mix of Bayesian Recursion and Inverse Bayes as a means to automate personalization. The goals of these papers is to outline a) why and how this is theoretically possible, b) whether it generates accurate predictions of student behavior and c) how this relates to students' sense of certainty or, as it is called with the Bayesian framework, subjective probability.

Theoretical Considerations for Automated Personalization through the Inverse Bayes Formula

Abstract

Individualization has been a goal for the educational enterprise for at least 2,500 years (Regian, Shute, & Shute, 2013, p.2). Renewed interest has been stoked by mobile computing, the Internet and recent increases in computer processing power. From Skinner's "teaching machines" to Khan Academy, the idea that technology can emulate an individual teacher for each student has inspired generations of educationalists and inventors. The following paper discusses the theoretical approach underlying an algorithm for utilizing subjective probability as a form of within-student variation. It outlines a model based on the Inverse Bayes Formula that generates estimates of constructs proposed by Snow: the Aptitude Complex and Situational Factors.

The model is systematically built from a basis in the theory of subjective probability, to utilizing sources of variation, to Bayes Theorem, then relaxing the assumptions of Bayesian rationality and inversion of Bayes Theorem, and finally a heuristic solution to estimating the range of the prior probability. The final model presented is the Inverse Bayes Filter (IBFi) and its utility for efforts to automate personalization is discussed.

Keywords: subjective probability, Bayes Theorem, Inverse Bayes Formula, Intelligent Tutors, assessment

Theoretical Considerations for Automated Personalization through the Inverse Bayes Formula

Individualization

Background. Individualized instruction has been a long-term goal of the educational enterprise. In 1899 Charles Eliot, President of Harvard University, bemoaned that, "Uniformity is the curse of American schools...Individual instruction is the new ideal" (Grittner, 1975, p. 325). Over a century later, the Secretary of Education Arne Duncan complained, "We need to take classroom learning beyond a one-size-fits-all model and bring it into the 21st century" (Department of Education, 2010b).

The longevity of personalization might be partly attributed to the malleability of the term individualization and the related terms personalization and differentiation. To Eliot's mind individualization referred to what would today be called elective classes, a now common feature of high schools, but the meaning of *individualize* has grown and been re-imagined far beyond this application. So much so, that The Department of Education felt the need to officially clarify the definitions of individualization and its sister terms differentiation and personalization:

- Individualization is pacing that caters to individual students
- Differentiation is altering instructional methods to cater to individual students' needs
- Personalization is the combination of individualization and differentiation but also including student choice in the mix

(Department of Education, 2010a)

As the definitions of individualization/differentiation/personalization have shifted one aspect seems to have remained though: the notion that technology can serve as a mechanism to enable students to receive the most appropriate content to their needs. In the US the idea that a machine could become a high fidelity replacement for the private tutor has been a part of the individualization debate from the beginning. The first patents for "personal education devices" appear in 1809, and Sidney Pressey is credited with the first published research on the use of "teaching machines" to pace set and provide individual feedback with respect to answer accuracy in 1926 (Benjamin, 1988, p. 705). The promises made on behalf of these technologies were not insubstantial: B.F. Skinner predicted that teaching machines would double the amount of information students could absorb and retain (Seidensticker, 2006, p. 103).

Over the 20th century three major strands of theory have contributed to the realization of technological individualization: Mastery Learning, Intelligent Tutoring Systems (ITS) and Aptitude-Treatment Interactions (ATIs) (Regian, Shute, & Shute, 2013, p.2). Mastery Learning is a systematic, teacher driven approach that groups students based on regular diagnostic assessments and adjusts instructional time so that every student in a class can meet tightly defined educational objectives (Bloom, 1968; Keller, 1974).

Intelligent Tutors grew out of the field of Computer Aided Instruction and incorporated ideas from Mastery Learning, such as tightly defined skills and using a human tutor utilizing a Mastery Learning process as an essential point of comparison (Desmarais & Baker, 2012). The goal of the Intelligent Tutor is to automatically assign an appropriate intervention to a student given their performance, although the lion’s share of research involves how to assign the right intervention to struggling students there is also research into how students who have mastered a skill should be treated (Kelly & Tangney, 2002).

Individualization in this theoretical frame is achieved through complex models of learner behavior that can allow the differentiation of student states. For example, understanding the difference between a student who mistakes multiplication for addition and the student who does not know what the multiplication symbol means.

The third strand, Aptitude Treatment Interactions grew out of psychometric research in the mid-1950s. It began when Cronbach set out to find "for each individual the treatment to which he can most easily adapt" (1957). And from this sprang a research program that lasted 25 years. ATIs seek to characterize interactions between student traits, most commonly general ability, and classroom condition variables. They differ from both Mastery and Electronic Tutors in that time is held constant and variation in ability is measured. This research program has commanded considerable time and resources but this effort failed to translate into insight that can be used in classrooms.

There is something intuitive behind all attempts at individualization, that if you can cater to the individual's needs, teaching will be more effective. ATIs formalized this idea in terms of aptitude, defining the individual's needs in terms of where they lie on a test distribution – low scoring students must have different needs to high scoring students. The extension of this argument is that, if low students and high students were no different in classroom relevant ways then there is no point to organized schooling. Yet a failure to consistently observe low-scoring students show improvement under conditions that correlated with their score indicates there is something wrong with this intuitive understanding. The likely flaw in the ATI strategy is defining difference in terms of relative student score.

Measuring Individuals. Despite the shortcomings of ATIs from a measurement perspective, and although ATIs did not ultimately prove their usefulness within mainstream educational practice (Speece, 1990), individualization has been dominated by the work of Cronbach and Snow (Cronbach & Snow, 1981). Cronbach and Snow laid out a framework that has impacted the broader enterprise of individualization in statistical terms and Snow in particular is credited with theoretical advances that help make sense of individualized classroom interventions (Shavelson et al., 2002). Specifically, Snow broadly expanded both the idea of aptitude (Snow, 1992) and the number of variables that ought to be considered when attempting to measure students for the purposes of individualization (Sinatra et al., 2001).

For Snow, aptitude was not a limited, general intelligence but rather a plethora of possible combinations of cognitive, conative and affective affordances, some innate and some learned, that could be applied to a given situation. These aptitudes form what Snow terms *the aptitude complex* and it is the fit between the aptitude complex and the environment that determine the success of a student at a given task (Snow & Lohman, 1984).

For example, a simple shot in a game of pool may match the skill level of a player well; she understands the physics of the game, she feels calm and she possesses adequate coordination. In this scenario, when situation and aptitude meet, the result is the successful execution of the shot (Figure 1A). A more complex shot, however, may be a poor match her understanding, coordination and feelings and so when aptitude and situation meet she fails to make the shot (Figure 1B).

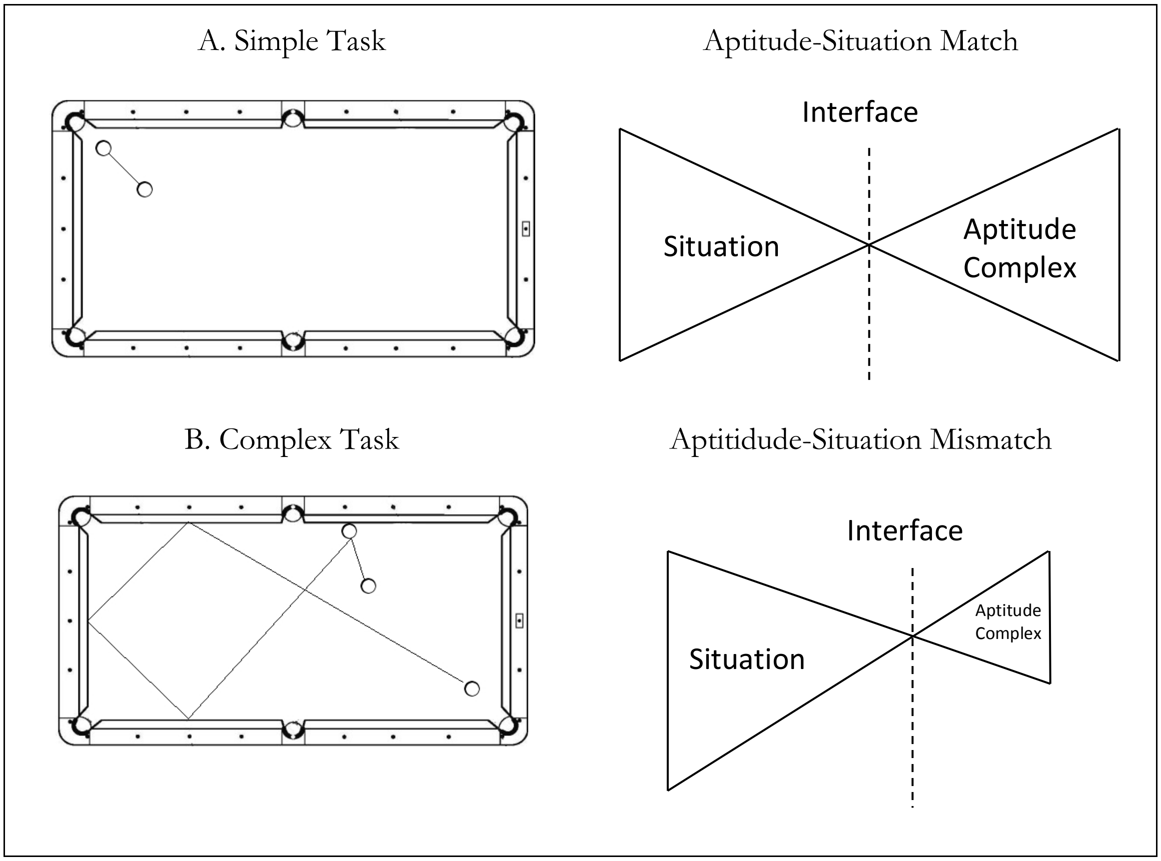


Figure 1. The match between aptitude complex and situation as represented by a game of pool with less and more complex pool shots.

Snow’s conception of aptitude is also dynamic, with both the aptitude complex and environment constantly in flux and a student’s performance unfolding “at the interface of person and situation” over time (Shavelson et al., 2002, p. 78). Teaching and learning therefore involved the reorganization of the aptitude complex to match particular environments and increase successful completion of tasks. In other words, according to Snow’s theory, building rich, individual profiles of students that describe both their aptitude and its interaction with situational factors could allow for effective individualized instruction. Snow saw the need for multivariate models that could achieve such a feat in an automated fashion though. He believed that the complexity of creating such individual profiles was beyond human capacity. For over a decade he worked on models derived from intelligence tests to do this (Yeh, 2012). Although these models were ultimately unsuccessful, the theoretical framework of viewing student performance as a match between aptitude complex and situation survives and informs several strands of research into individualization, such as Performance Factor Analysis (Pavlik, Cen, & Koedinger, 2009) and the bulk of psychological research into individual differences (Jonassen & Grabowski, 2012).

The aim of this paper is to use Snow's theoretical system of balance between aptitude complex and situational factors to build a mathematical model of individual student learning. This model, if reliable, could advance the aim of automated individualization, to provide the most appropriate conditions to a each student. This will bring together three main pieces: 1) Utilization of subjective probability as a source of within-student variation to avoid reliance on between-student variation as ATIs do 2) the perspective of rational Bayesian models to provide a means of processing the subjective probability and 3) the inversion of these Bayes models to generate parameters for each student over time.

Model Design

Choosing a Source of Variation

Aptitude Treatment Interactions were based on models of intelligence and therefore took as their essential measurement the variation *between* students on intelligence-type tests. The nature of the variation that they preference is important as using only between-student variation likely had a lot to do with the failure of the models to create reliable results (Speece, 1990). The strategy ATIs use to make inferences is to look for a relationship between a student's relative test score (aptitude complex) and some environmental variables (situational factors). If groups of students differs with respect to this relationship then this represents a point of individualization. Figure 1 demonstrates what this looks like when test scores and number of homework hours show an interaction. The problem with this strategy, however, seems to be its reliance on between-student variation. The heterogeneity of both individuals and environments is far greater than the model can reasonably account for and interactions tended not to be replicable and interventions based on the ATIs ineffective (Speece, 1990).

In contrast to ATIs, the two other domains that have contributed substantially to automated individualization, Mastery Teaching and Intelligent Tutors, look to preference a different form of variation: variation across time. This is not to say that ATIs did not incorporate time, but rather they start with variation between students and then scale across time – variation between students is the reference point from which other factors were measured. Conversely, mastery learning looks to group students according to their growth towards skill proficiency, based on skill based assessments and intelligent tutors tend to validate their measures through forecasting accuracy (Mäkitalo-Siegl & Fischer, 2011).

Validating individualized measures. The varying approaches to time between ATIs and Intelligent Tutors have important consequences for how the two fields approach validation. ATIs require time to be constant for all students so that the variation between students over this window can be compared. The comparison and validation of ATIs was developed from IQ research and as such utilizes methods of validation common in the psychological sciences. They involve Frequentist significance testing to determine whether a relationship exists between variables. Validity of an interaction means that the relationship is likely to exist within an acceptable tolerance of possibility, as determined by comparison to an appropriate distribution (Messick, 1995). In contrast, Intelligent Tutoring Systems, such as those utilizing Bayesian Knowledge Tracing, use validation more in common with forecasting methodology. Since students move at their own pace, there is no standard time variable, so validation is based not on the existence of a significant relationship, but rather on successful prediction of future student performance.

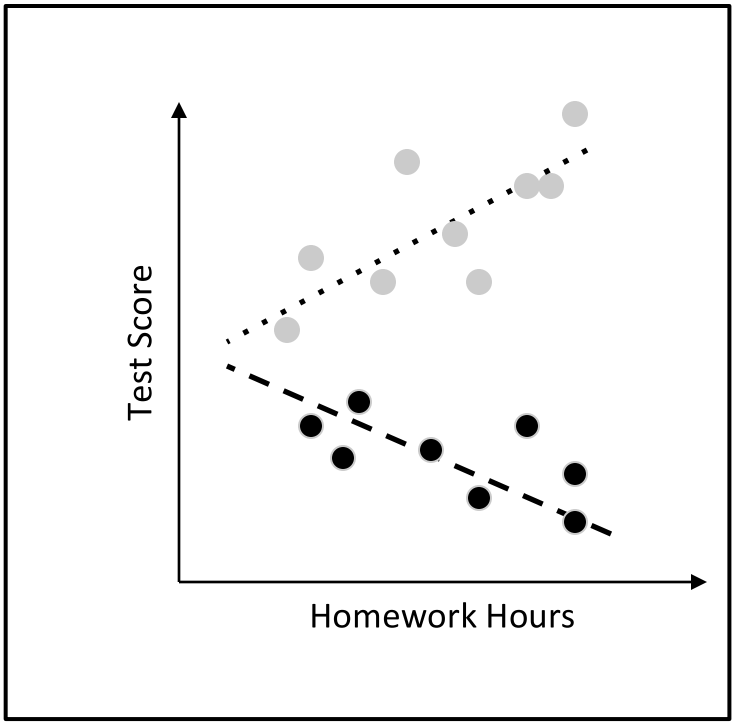


Figure 1. An Aptitude Treatment Interaction. Gray and black points represent students whose test scores show the opposite interaction with number of homework hours performed.

Parameterizing individualization. Despite having the express aim of individualization, Intelligent Tutor, Mastery and ATI models rarely use individual level parameters – parameters that exploit variation within individuals over time. ATIs utilize between-student variation almost exclusively to group students relative to their peers (Cronbach & Snow, 1981). The major Intelligent Tutor model, Bayesian Knowledge Tracing (BKT), mostly utilizes skill level parameters despite its original conception including individual level parameters (Corbett & Anderson, 1994; Yudelson, Koedinger, & Gordon, 2013, p. 1). An alternative Intelligent Tutor model, Performance Factor Analysis, produces skill and item level parameters utilizing between-student, between-skill and between-item variation (Pavlik et al., 2009). Mastery Learning methods also group students based on between-student scores on assessments that define skill-based variation (Black & Wiliam, 2006).

Practicality has likely had a substantial influence in the choice to model between-student change instead of change in individuals over time. Collecting individual level data has historically been complex and expensive (Klein, Dansereau, & Hall, 1994). Collecting, scoring and storing the information on many students was relatively unrealistic prior to the introduction of cheap computer storage and processing capabilities. As such, sample-based statistics were designed to deal with this problem by allowing inference from a limited number of samples. Sampling methods remain the dominant way of dealing with educational data problems (Ary, Jacobs, Sorensen, & Walker, 2013, p. 169).

The preference of other forms of variation has also meant that individual level parameters tend to be added to models that predict variation between students or skills, rather than the other way around. The tendency for individual level parameters to play second fiddle exists even within experimental psychological models of dynamic systems, in which individual level time parameters are mixed into models of mean differences between individuals (Borsboom, Kievit, Cervone, & Hood, 2009).

The failure of ATIs to predict future performance and the limited success of BKT to improve predictions of future student performance with individual level parameters may reflect that these sources of variation do not reflect the same underlying processes (Borsboom, 2005). There is therefore some interest in building models that use as their base individual change over time. These models use an approach in which each student is given an individual prediction of their performance, separate from every other student. The question remains however, how do we accomplish this?

The key requirements of such a model would be:

- Based on within-student variation
- Forecasting accuracy used to validate the model for each student

The aim of the following section will be to use Snow's aptitude complex and situational factors to predict student performance using these model requirements.

Operationalizing Theory

Modeling Aptitude Complexes and Situational Factors

The approach taken here to achieve the desired model is to formalize Snow's construction of aptitude complexes and situational factors in terms of subjective probability. Subjective probability has been intermittently proposed as a way of measuring student performance but it has often been associated with the idea of the "rational actor", a constraint that has made it impractical in the educational domain. New technologies and faster computer processing speeds may make it worth revisiting though.

Subjective probability. "Statistics is the study of uncertainty" (Lindley, 2000, p.294), but for the most part it is the study of the uncertainty of an observer of an event. In the case of assessment it is the uncertainty of the psychometrician with respect to the measurement of an observed psychological construct in a test taker. In this case the uncertainty *belongs* to the psychometrician, but is there value in modeling the uncertainty of the student?

By instead bringing a calculus of probability to bear, not on the psychomterician's uncertainty, but on the uncertainty of the student, there is the possibility that a useful source of within-person variability could be unlocked. We might then build a model based on within-student variation that may not suffer the same limitations that ATIs did with their focus on between-student variation. In other words, we will derive probability-based measurements from the perspective of the student. Certainty in educational measurement is not a new idea and was advanced within the educational context by Coombs and then de Finetti in the 1950s and 60s:

> A different outlook appears, sometimes rather stealthily and marginally, sometimes clearly and firmly expressed, when the existence of partial information, or knowledge, is anyways considered. Indirect ways are those introducing notions like the confidence (of a subject in his choice), the "standard of assurance", the "sensitivity" of a choice, and so on. The direct way is to put forward partial knowledge as the very essential basis for the whole theory: that has been done in the fullest manner (as far as I know) by Coombs (1964), but it seems possible and necessary, in my opinion, to take one more step forward. This step consists in the straightforward interpretation of any partial knowledge in terms of probabilities (of subjective probabilities, to be precise; such distinction is in the subjectivist theory of probability, to what I adhere, for any probability is there but a subjective belief). (de Finetti, 1965, p. 87).

In 1963-6 de Finetti set out to consolidate ideas in Subjective Probability with those in educational assessment. The growth of the multiple-choice examination seemed to him to be the natural testing ground for his theories about probabilistic reasoning and mathematical psychology. However, the educational research community, including the then leaders of assessment Lord and Novick (1968), mostly rejected the resulting theories as being too easy for students to manipulate. Instead, de Finetti's work found a home in the world of business, political science and economics where his ideas contributed to decision theory, game theory, and utility theory (Schlaifer & Raiffa, 1961).

The failure of de Finetti's work to gain traction within education, while it blossomed elsewhere, is telling of the differences between fields. The first two pages of his treatise on the application of subjective probability to education detail how students need to be trained in order for the assessment to work. Students must understand probability theory, understand the scoring method, and want to gain the highest score possible. In short, they must be informed and rational actors. The idea that students would act within this limited sense of rationality was a stretch that educationalists were not prepared to entertain (Sullivan, 2006). In fact, the irrationality of student decision-making has been well documented, with everything from choices in the cafeteria (Gottfries & Hylton, 1987) to whether to cheat on an exam (Tibbetts, 1997) being used to demonstrate students' failure to comply with definitions of rationality. Within education, rational models are most commonly used to explain how students choose which college to attend. Yet, even this substantial literature is critiqued for having unrealistic assumptions, such as the assumption that students are adequately informed about college (Des Jardins & Toutkoushian, 2005, p.194). This assumption seems to be the point of distinction between education and other fields such as economics. Whereas it is reasonable to assume that a consumer or politician is adequately informed, education, by definition, implies that information is lacking that needs to be learned. It is not possible to have a fully informed student to model, as this would defeat the purpose of the educational process. However, the idea that businesses, consumers, managers and politicians can be fully informed is, at least, definitionally possible (Gupta, 1994). In the following section I will outline why this difference between education and other fields is important, and why the idea of subjective probabilities in educational assessment should be revisited.

Connecting subjective probability and Bayes Theorem. Subjective probability is an extension of probability theory that accounts for the individual's sense of certainty, often characterized as a *partial belief*. It is sometimes used to refer to personal opinion or expertise, but is strictly considered a probability only in cases where it adheres to a set of logical statements, the two most common of which are Kolmogorov's Axioms or Cox's Theorem. As such it is referred to as a logical, as opposed to classical or empirical, probability. These more familiar forms of probability are defined in terms of proportion (the probability of rolling a 6 on a die is 1/6). However, subjective probability has some useful properties that distinguish it from these objective probabilities. Chiefly, subjective probabilities can be true for singular events such as "the probability it will rain tomorrow is 70%" (tomorrow will only happen once), as well as repeated events, "The probability that the coin is heads is 0.5" (the coin can be flipped many times).

Subjective probability might simply be a curiosity without Bayes Theorem, the relationship between conditional probabilities devised by the 18th century minister Thomas Bayes. Originally it was intended to solve the trivial problem of guessing the proportion of black and white marbles in an urn, but it has come to be the functional workhorse of subjective probability. When formulated with respect to a hypothesis and available data Bayes Theorem allows reasoning about uncertainties based on prior knowledge and incoming data in accordance with:

$$P(hypothesis|data) \propto P(data|hypothesis) \times P(hypothesis) \qquad (1)$$

An example from Pierce (1877):

A newborn baby sees the sun rise and then set but is unsure whether the sun will rise tomorrow. What is a reasonable probability estimate for the baby to have of the sun rising in the morning? At that point her experience with sunrises, the prior *P(sun rising)*, is very limited, the probability of the sun rising is equal to it not rising, but the data she has available, the likelihood *P(sun rose today|sun rising)*, is overwhelmingly in favor of the sun rising. Bayes theorem provides a best estimate of the posterior probability, *P(sun rising|sun rose today)*, or the probability of the event, given the available data.

In sum, Bayes Theorem gets it strength from the ability to systematically incorporate new data as it becomes available, a process known as Bayesian updating. In the above example, as the baby witnesses more sunrises, the value for the posterior probability can be exchanged for the prior, the new data are incorporated into the likelihood, and a new estimate of the probability of the sun rising given her experience with sunrises can be calculated.

Subjective probabilities in education. To de Finetti, educational assessment seemed to lend itself to subjective probabilities rather than classical probabilities. To his mind, it is unclear what the proportion applies to in the case of a single student, answering a single question (Borsboom, 2005, p. 74). It may variously mean the proportion of students who answer a question correctly and/or the proportion of times a student gives the same answer to that question. The former sense is problematic insofar as the proportion of students must have a substantive meaning that is dependent on all the students. For example, a correct answer to a question could mean that a student either understands a concept and an incorrect answer means that a student does not understand that concept. If a correct answer means something different for a subset of students though, such as that they can infer the correct answer from their understanding of a different concept, the meaning of the proportion becomes unclear. Likewise, the proportion-over-time interpretation is problematic to operationalize. Besides the drudgery of requiring the same question to be asked multiple times, the nature of learning is such that it will, by definition, change the proportion over time. As students learn, they will change from having a proportion of zero correct to 100 percent correct and the meaning of the proportion will change as they do. For example, a student may be incorrect on the first three trials of a question but they will learn the answer by the fourth, their proportion will be 0.25 over these trials but from this point forward they will generate the correct answer so their proportion should be close to 1.

However, subjective probabilities, with their interpretation as “strengths of belief”, avoid the constraints of classical probabilities described above. They can be true for a single event and change as conditions change. To de Finetti, such properties seemed to fit the longitudinal nature of education and by using them he could avoid referring to proportions at all, instead using students’ own subjective probabilities – their confidence in their answers – to feed the model.

Confidence Measurement. Interest in using student confidence as an assessment measure arose out of the mathematical formalization of subjective probability at the end of the 19th century (Estes, 1976, p.37). Since 1913 researchers have sought to apply this body of theory to educational assessments (Woodworth, 1915, p. 10). The initial motivation from the educationalists' perspective was to determine if querying student confidence could provide useful additional information about student performance (Echternacht, 1972). Over the last century the utility of confidence testing has been demonstrated in terms of test reliability (Ebel, 1965; Rippey, 1968), identifying guessing (Taylor & Gardner, 1999), separating students based on their level of understanding (Gardner-Medwin, 1995), increasing student understanding (Echternacht, 1972; Gardner-Medwin & Gahan, 2003; Ramsey, Ramsey, & Barnes, 1987) and explaining answer changing (Skinner, 1983). Yet, despite continued interest and positive reviews of the method, empirical research concerning confidence and educational assessment seems to be perpetually idling as an interesting idea, without widespread implementation or research interest.

This reluctance may in part be due to the criteria, set down by de Finetti, for confidence to be considered a legitimate psychological measurement technique: that a scoring system could be devised that the student could not *game* to her advantage, and that the measurement unambiguously improved reliability and validity. With respect to the first challenge, although many systems were developed, none have become preeminent, and with respect to the second, Lord and Novick (1968, Chapter 16) dismissed any improvement as unlikely, dealing a death-blow to the method in the eyes of many psychometricians (Echternacht, 1972).

Despite these dismissals, approximately every ten years researchers suggest using some measure of student confidence as a viable option to mitigate problems such as guessing, yet little expansion beyond a handful of scoring techniques has been achieved (Frary, 1989, p.88; Gardner-Medwin & Gahan, 2003, p.152). The paucity of study of confidence in education is in contrast to its extensive investigation in psychology, mathematics, statistics and business (Eser, Holbrook, & Colbert, 2012, p.28).

Bayesian rationality. There could be any number of reasons why education has not yielded more research into subjective probability and student confidence, or wider implementation within national or state assessment systems. For a long time the extra logistical burden of collecting and analyzing the data was not feasible (Taylor & Gardner, 1999, p. 355). But technological advances in computing and data storage have reduced those burdens considerably over the last 30 years. Beyond logistical obstacles there is a lack of theory about confidence that is endogenous to education unlike economics (decision theory), management (recursive Bayesian updating), engineering (High Confidence Theory and artificial intelligence models), mathematics and statistics (confidence intervals, subjective probability, Bayes) in which substantial theories have been developed around confidence.

What these theories have in common is that they utilize Bayesian methods and accept the constraints that these methods put on inference. With respect to modeling how students perform these constraints are presented as the assumption of Bayesian rationality. de Finetti outlined the key pieces of this rationality in three points:

1. The student must divulge all her possible hypotheses about a topic. For example, a multiple-choice question must have options that reflect all possible contingencies that a student may think of by including an "other" option.
2. The student must understand probability theory and the scoring method used so that she is able to optimize her score
3. The student must want to optimize her score

These assumptions are difficult to maintain – the need to test both the understanding of the test and the motivations of the students is a substantial barrier. Further, and the value of the information gained for that effort is unclear, although the idea of an ideal Bayesian learner as a point of comparison has not found utility for students outside of career decision-making

(Gordon, 2003; Harren, 1979) and risk taking behavior (Tibbetts, 1997). The models demonstrate an ideal version of decision-making, yet there are few instances when such a reference point is sought in educational assessment. Students are considered to be imperfect decision makers for many reasons and it is unclear what benefit is gained by comparing them to an ideal (Leclercq, 1993). Subjective probability though does seem to have some value in an education system in which personalization is a goal – all students possess their own personal record of belief. If this belief network could be accessed then it could be a useful, dynamic source of information about learning that could be used to tailor instruction.

Bayesian rationality is not product of subjective probability or even Bayes Theorem itself, but rather the way these ideas are operationalized. The prototypical Bayesian learning experiment compares a population of decision makers with the ideal Bayesian. A recent example is the work of Gopnik et al. in which children are tasked with learning species of imaginary animals (2004). The children are given example pictures of the animals and then told to name unidentified examples. On average, as the number of example animals increases, so does the child's ability to recognize them, and the shape of the curve on average across many children is in accordance with the curve generated by Bayes Theorem. Results such as this are taken as evidence that people use Bayesian reasoning to learn, but there are of course many studies that show how on average people do not comply with Bayes Theorem. Yet both these arguments miss the point.

The idea that there is an average subjective probability takes the wind out of the sails of the subjectivist, who claims that subjective probability can differ for everyone. The presumption of the average subjective probability is not built into the model, but into the way the model is set up. Rationality is not only dependent on the relationship, but on how the data and hypotheses are established. This establishment is done in two ways:

1. The presumption that you can characterize the likelihood as the available data defined by the researcher. EG – if I tell you some key information you will rationally incorporate that into your view of the problem.
2. The presumption that you will update your beliefs in an efficient way. That the posterior will be transferred to the prior completely and without delay.

These strategies have shown great utility in calculating the certainty in non-psychological events, but are restrictive in the case of educational assessment. However, the assumptions can be relaxed and so the model can adapt to fit the nature of assessment. It is this adjustment that educational scholars have not previously looked into and that may provide a unique pathway to leverage Bayes Theorem for assessment purposes.

Relaxing the Assumption of Rationality (Inverse Bayes). As it stands, Bayes Theorem does not dictate how the data are incorporated into the likelihood, nor does it endorse Bayesian updating. In fact, Bayesian updating has been discredited as a formal proof of the Subjective view of probability (Douven, 1999). Bayes Theorem only states that there is a relationship between the posterior, likelihood and prior and a Bayesian reasoner will calculate their posterior in proportion to the likelihood of the data available, multiplied by their prior knowledge. There is no stipulation that all people in a population hold the same prior information, nor that they refer to the same data. Overall, it does not stipulate in which direction the equation should be calculated.

It is therefore possible to use Bayes Theorem without the updating feature and without incorporating data into the likelihood, but it requires reversing the configuration it is usually presented in. Instead of calculating a posterior, we would begin with posterior results and ask what sort of prior and likelihood would have led to those results. This form, called Inverse Bayes Formula, is a surprisingly recent re-interpretation of Bayes Theorem and it has yet to be applied within the social sciences (Tian & Tan, 2003). It has however been well established within distributional science, and utilized with missing data problems (Tian, Tan, & Ng, 2007), optimal control theory (Friston, 2011), and medicine (Matthews, 2001).

If we use the Inverse Bayes Formula to dissect subjective probability we are no longer describing rationality, but rather whatever flawed way someone reached his or her decision. We are relating how they weighted their prior knowledge against whatever data they thought relevant. This weighting is particularly useful to educational assessment as it provides a description of what a student knows and how their situation is influencing their decisions.

Model Building

We now need to bring ideas from Inverse Bayes and subjective probability together within Snow's relationship between the aptitude complex and the situational factors. Once we have characterized behavior in terms of probability we can start to model how conditions and student characteristics come together to produce student performance.

Operationalizing Probability

Lock & key analogy. Consider a collection of six locks and three keys. Key 'A' opens one lock, key 'B' opens two locks and key 'C' opens three locks (Figure 2).

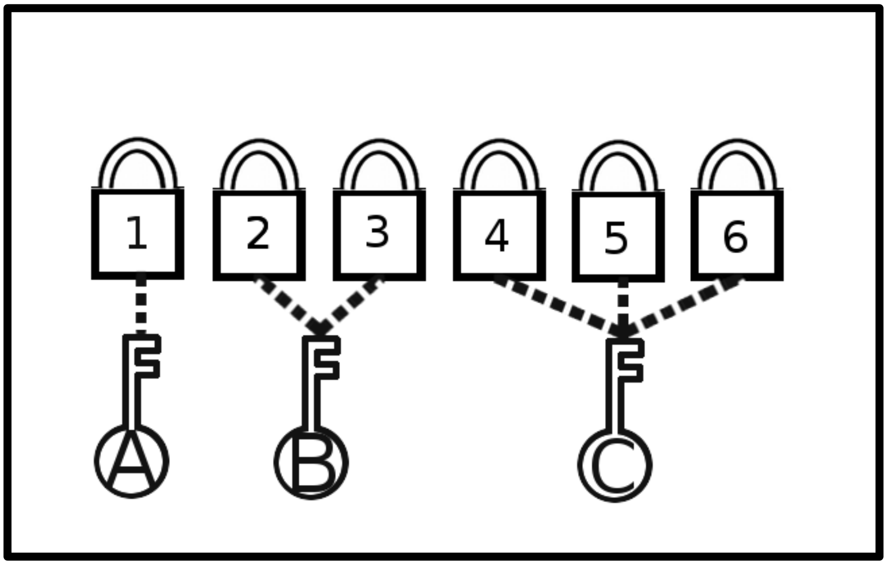


Figure 2. – Lock and key analogy of tasks and strategies.

The locks represent tasks and the keys represent successful strategies to complete those tasks – a successful strategy *opens the lock* of a task. We might then consider that a curriculum is made up of successful strategies, and that the reason that a strategy is included in a curriculum is that it opens many locks (it is more like key 'C' than key 'A'). In other words, there is a generality to the strategy. For example, factorization is not taught so that students can solve a single equation; likewise, vocabulary is not used in a single sentence. Even facts, such as the capital of China is Beijing, hold generality and stability. Beijing may be relevant to a task in economics or in history, but regardless it remains the capital of China.

Extending this analogy further, we might consider that learning in a formalized environment is a process whereby students construct copies of the curriculum keys. Assessment is then a record of the application of those keys across different locks. Summative assessment is a record of which keys a student has, while formative assessment is a record of the process by which the keys were constructed. In either case the central question is *how to determine that a student has a fidelious copy of a key*, or in other words, *how similar is a student's copy to the key defined by the curriculum?*

The basic answer to this question is that if a student successfully completes a task then she has an accurate copy of a key. This is an unsatisfactory characterization though because it gives no indication of generality or stability. For a test item, a lack of generality of stability may mean any number of possible alternate explanations such as the student guessed the correct answer (neither general nor stable), or has memorized the answer to that particular question only (stable but not general).

Probabilistic interpretation. Instead of considering a correct answer as evidence for an accurate key we might instead consider that both locks (tasks) and keys (strategies) are probabilistic. We can do this at the level of an individual student answering an individual question. Tasks are probabilistic in the sense that they are more or less common *in the experience of the student*. We might extend this perception of experience not only to tasks that have unambiguously happened in the student's life, but also tasks that they imagine can happen.

In contrast, strategies are probabilistic in the subjective sense of Cox (1946); a given strategy has a level of certainty associated with it. As such, a student can be more or less certain of a strategy. With these two probabilities defined it is possible to construct a joint probability distribution, where T_A is the task being undertaken in the present, and for simplicity's sake we will consider that all other tasks can be grouped into T_B and T_C. Again for simplicity let us also consider that the strategy either be true or false.

Table 1. Joint probability distribution of strategy and task

	Strategy		
	True	False	
Task$_A$	0.3	0.1	0.4
Task$_B$	0.2	0.2	0.4
Task$_C$	0.1	0.1	0.2
	0.6	0.4	1.0

It is worth noting that the values in the dark gray cells represent the student's mental state and are untestable to us. The marginal probabilities, conditioned on strategy (black cells) however are measurable and of interest. Of particular interest from an assessment perspective is the marginal probability of True, $p(strategy_{TRUE})$, 0.6. The marginal probability represents the certainty of the strategy being 'true' considering all the experience of the student. Intuitively this seems like a valuable piece of information, as it incorporates not only the task being undertaken but all the tasks, real or imagined, that a student is bringing to bear on a strategy in that moment. In terms of the lock and key analogy, it is a measure of how general the student considers their key to be.

Problematically, there is no brief way of determining all the joint probabilities between a strategy and experience. Asking students about all their experiences would be time consuming and likely change the marginal probabilities around those experiences, although Winkler (1967) had some success with this strategy. We can however easily calculate the marginal probability from a conditional probability in accordance with Bayes Theorem, provided we can define a posterior, $p(strategy_{TRUE} | task_A)$, and likelihood, $p(task_A | strategy_{TRUE})$:

$$P(strategy_{TRUE}|task_A) \propto P(task_A|strategy_{TRUE}) \times P(strategy_{TRUE}) \tag{2}$$

The posterior is the probability of the strategy being true, given the task. With respect to how we have defined the probabilities in question, it is the level of uncertainty given the commonality of the task in question, from the perspective of the student. Or in other words, the level of certainty the student has in her answer on a scale between 0-1 based on how often she has encountered these circumstances. For instance, a student can be 100% confident in her answer to an item, not confident at all, or any level in between. This is a similar approach to the Decision Theory of Schlaifer and Raiffa (1961) and the Cognitive Bayesian approach to reasoning of Griffiths (Perfors, Tenenbaum, Griffiths, & Xu, 2011).

The likelihood is the probability of the present task given the student's belief in the strategy being true. For example, if the student believes strongly that the answer is true, does the task and its circumstances support this belief. In other words, it is the degree to which the present task pushes the student away or towards the strategy being true. In this way the likelihood provides a mechanism through which the influence of the student's circumstances can be inferred. Do they help or hinder the student reaching the correct answer.

An Example. The fundamental idea behind applications of Bayes Theorem to people's thinking such as Decision Theory (Schlaifer & Raiffa, 1961) and Cognitive Bayes (Griffiths, Kemp, & Tenenbaum, 2008) is to change the vantage at which it is applied. For example, instead of conditioning on the situation from the perspective of a researcher or an assessor (e.g. – the probability of the student being correct given the item) we condition on the situation from the perspective of the person being assessed (e.g. – what is her hypothesis, and on what data is she conditioning). For example, if we were studying a student as she answers the following item:

Koalas are:

A. Carnivores

B. Omnivores

C. Herbivores

D. Calmivores

We could devise a model for the way she approaches each answer A, B, C & D:

$$P(koalas\ are\ herbivores|data) = \frac{P(koalas\ are\ herbivores)P(data|koalas\ are\ herbivores)}{P(data)}$$

In this model students weigh the likelihood of the data they have on hand against their prior beliefs, and as more data are presented, they are able to update those beliefs. For example, we might show a student pictures of koalas and every time we revealed a new picture, we asked the student whether they thought the koala was a herbivore. In Bayesian Updating we model the process of their opinion as a Bayesian process where each new picture was a datum that changed the likelihood, generated a posterior and then that posterior became the new prior. This application of Bayesian Updating underlies features of Decision Theory and Cognitive Bayes. Where it departs from Decision Theory and Cognitive Bayes, is over the efficiency of that updating mechanism.

The Decision Theorist will assume that updating is efficient or *rational* (Oaksford & Chater, 2007) and that there is error in the individual's reporting of their posterior. Decision Theoretic questions tend to be along the lines of “Do financial analysts make rational decisions about market conditions?” However, if we apply IBF the individual can state her own posterior probability accurately, but the incorporation of this new information is not necessarily performed efficiently. Data may not be attended to, or it may not be incorporated into a person’s beliefs. We can then ask the question, “How do the following conditions impact this individual’s prior probability in a specific task?”

Generalizing the Model. It is possible to expand the joint probability distribution above (Table 1) to reflect a continuous probability distribution so that True/False becomes a level of certainty between zero and one. This array can be represented as a 3-dimensional surface plot (Figure 3A), and the conditional probability of a strategy given a task is taking a single-point, horizontal slice across the plot (Figure 3B).

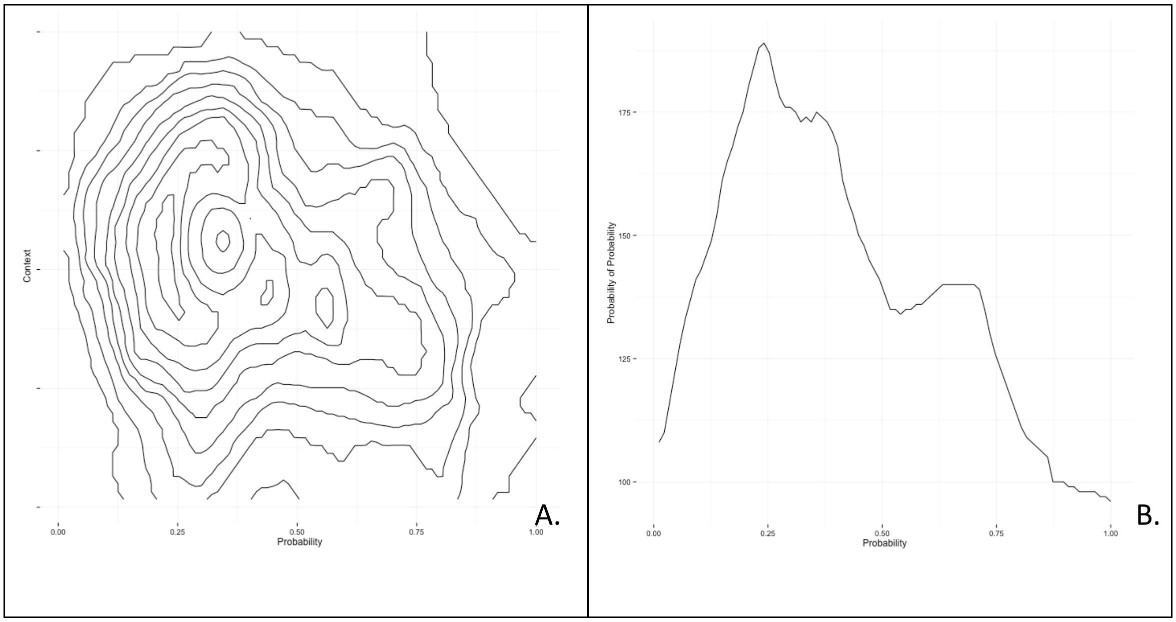


Figure 3. A contour map representing the probability density across many contexts (A) and a slice through the continuous distribution representing the prior or marginal probability according to the student. Here it has a peak at 0.25 and 0.60. (B).

For any given task/context we are unsure where the cut has been placed on this plot, but this is enough information to characterize the marginal probability, provided we can parameterize the posterior and likelihood. Effectively, this means that if we can measure the posterior and we can reasonably estimate the likelihood, we can generate a prior distribution. As if we had taken a one-context slice through the 3D surface plot to yield the prior distribution for a particular context (Figure 3B).

The remainder of this paper deals with how to convert a point estimate of the posterior probability into an estimate of the prior distribution and likelihood. The aim of this explanation is the construction of the Inverse Bayes Filter.

Parameterizing the Likelihood & Prior

Inverse Bayes Filter

In Snow's terms, the Inverse Bayes filter seeks to determine the relative contributions of context and the aptitude complex to student performance in accordance with the Inverse Bayes Formula (IBF). The aptitude complex is whatever cognitive, emotional and conative resources a student brings to a task. Contexts are the conditions of the task that impact a student's performance. For example, a student may be certain about her name, but within a high stress context she may not be able to report it. Likewise, she may be very uncertain about the laws of thermodynamics, but if we provide enough context cues she may be able to choose the correct answer from a selection.

Inverse Bayes determines how knowledge and context should be weighted for a student, given their answer according to logical probability. Bayes Theorem posits that the conditional probability of a hypothesis (posterior) is proportional to the product of the probability of that hypothesis (prior) and the likelihood of the available data conditioned on the hypothesis (likelihood):

$$P(hypothesis|data) \propto P(hypothesis) \times P(data|hypothesis) \qquad (3)$$

Bayes Theorem then gives the relationship between the aptitude complex and behavior according to context. The following graph demonstrates this by showing how, according to Bayes theorem, as the aptitude complex decreases (large dashed line), context (solid black line) must increase quickly (become very friendly) to increase performance (small dashed line). In other words, context can compensate for not knowing something to some extent, but there are limits to this relationship.

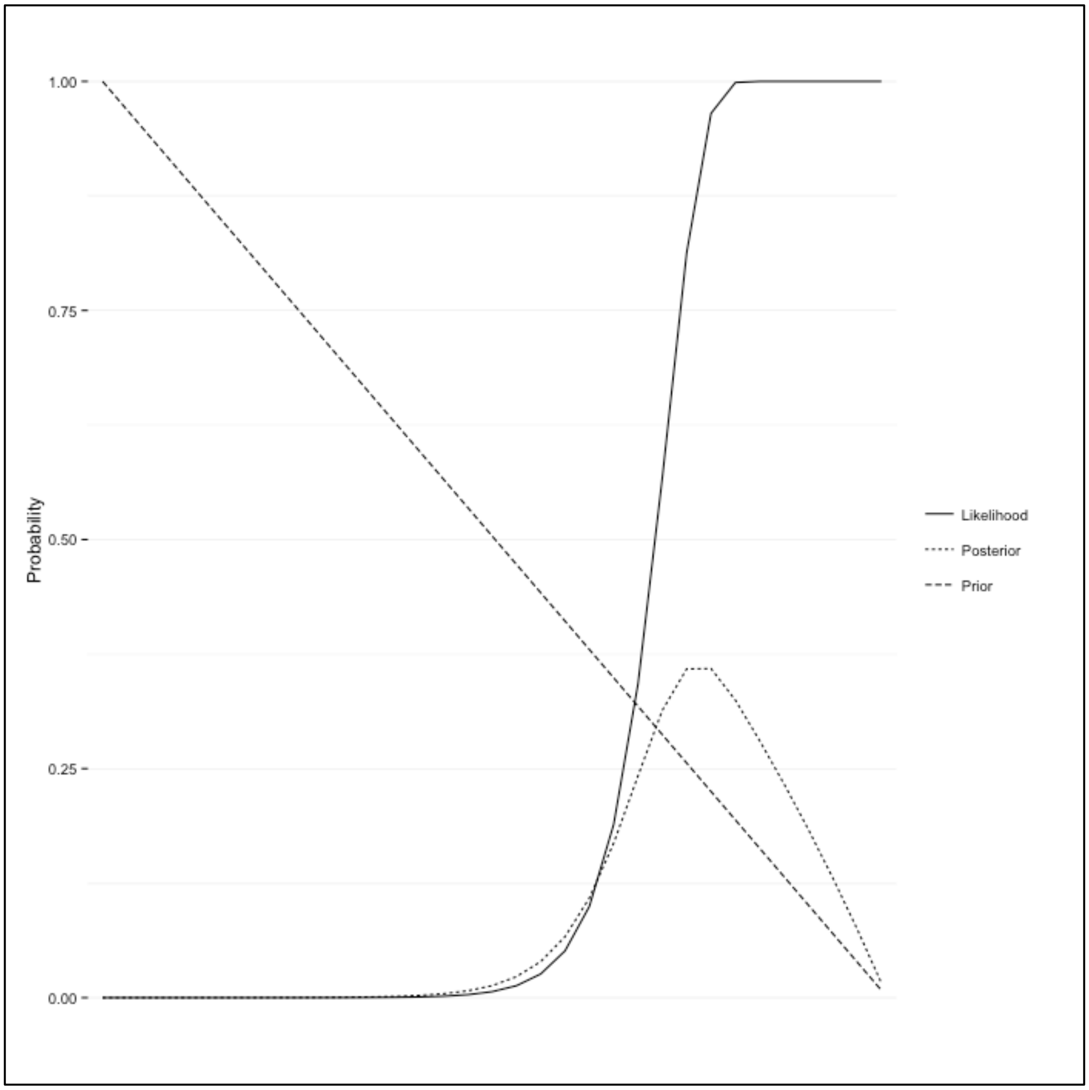


Figure 4. Theoretical values for performance, aptitude complex and context demonstrating the balance between knowledge and context. Performance can be high even if knowledge is low, provided there are enough context clues, likewise, partial credit can be low even if knowledge is high if the context is very unfavorable.

An inverse Bayesian approach to this problem differs from these previous examples though in that it does not treat the posterior as a stand in for a stimulus. For example, if an experimenter randomly assigned one group of students tablet computers, a Bayesian approach would treat tablet computers as a stimulus that had been equally applied to each student and differences in their performance would on average reflect the impact of the device. Differences between students in this model are treated as random error. Conversely, in an Inverse Bayes model we try to characterize these differences within the posterior, and they are interpreted as how each individual student is interpreting their context. In other words, in a Bayesian model all students have the same posterior and in an Inverse Bayesian model the posterior is free to change for each student independently.

Estimating the prior and likelihood. There are an infinite number of possible prior and likelihood combinations for each posterior. However, the range of the likelihood will differ dependent on the posterior, but the range of the prior will always be between 0 and 1 (Figure 3). That the posterior gives no information about the range of the prior is an issue if we are using the posterior to estimate the prior. To deal with this issue we can implement a heuristic that assumes some properties of student certainty.

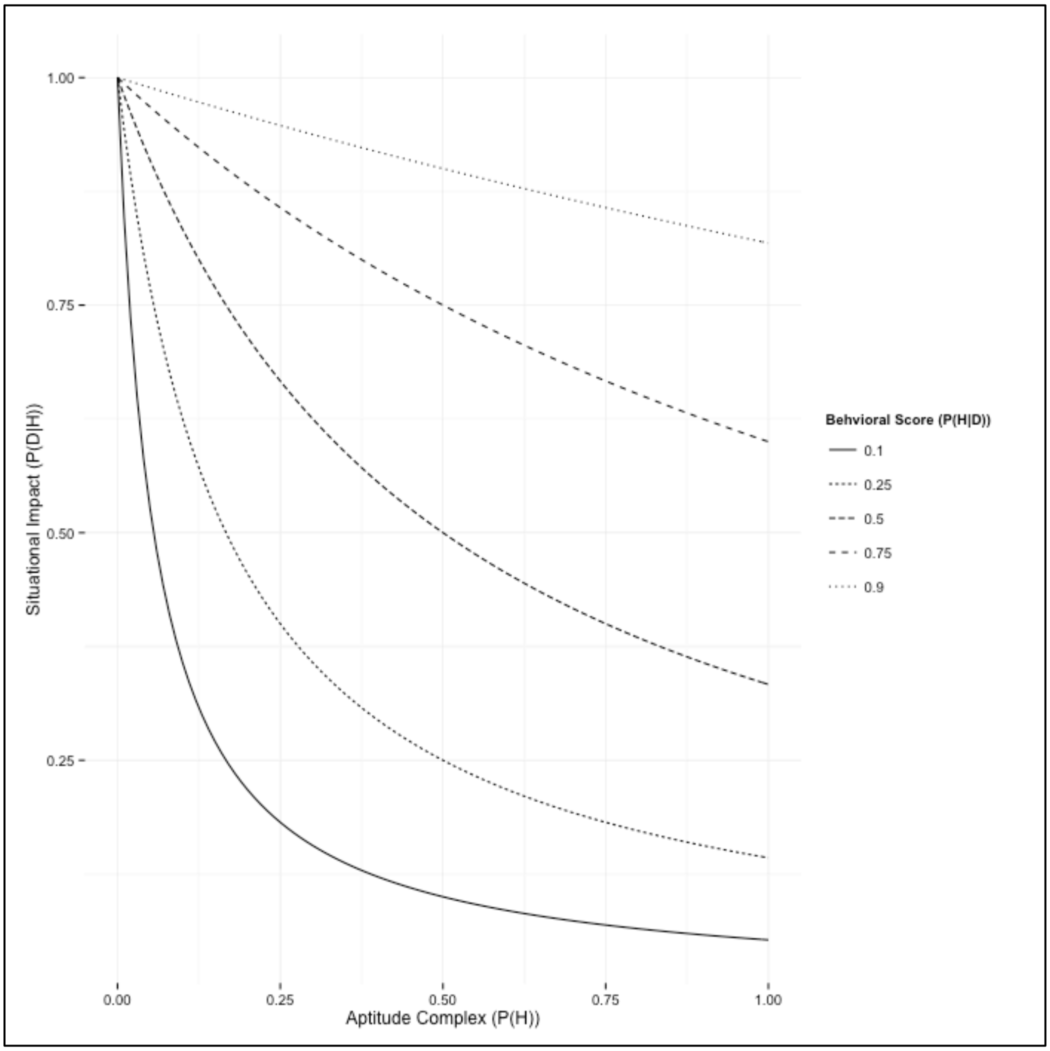


Figure 5. The range of likelihoods (contextual impact) vs. priors (internal) for different levels of posterior. The range of the likelihood changes dependent on posterior but the range of the prior is always 0 – 1.

Heuristic for estimating the prior probability. To differentiate the range of the prior for each possible posterior it is necessary to adjust the formula to account for how the range of the prior might change dependent on the posterior value. One way of doing this is to assume that certainty has lower variance at extreme values. When students are very uncertain, then they at least know they are uncertain. However, if they are somewhat uncertain they could be certain or uncertain. We can do this by describing the variation in certainty as a Gaussian distribution where the variance of the distribution is a quadratic based on the value of the mean of the Gaussian (Fig.4):

$$f(x,\mu,\sigma) = \frac{1}{-(\mu^2+(\mu-\left(\frac{\mu^2}{2}\right))\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2(-(\mu^2+(\mu-\left(\frac{\mu^2}{2}\right))^2}}$$

(4)

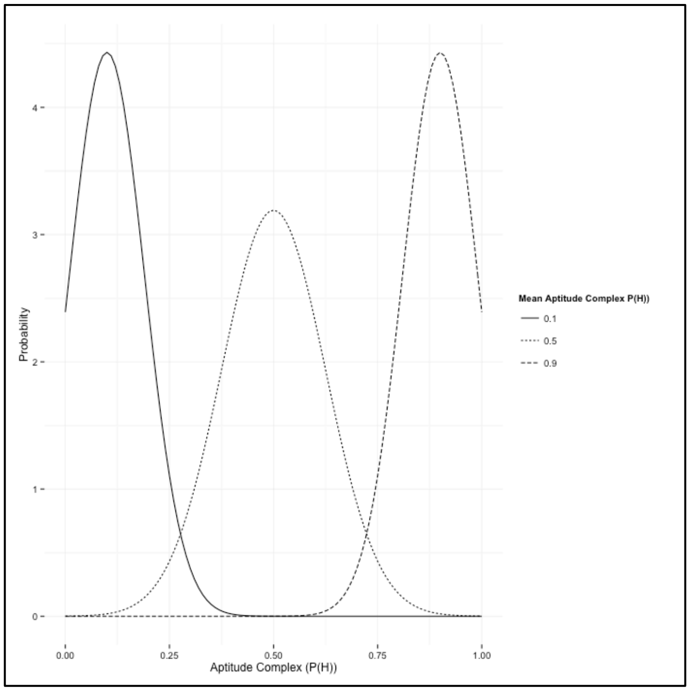


Figure 6. Gaussian curves describing the variance of p(h) or *certainty in certainty* at certainty levels of 0.1, 0.5 and 0.9.

By using this heuristic we can produce a limited range for the prior based on individual student posteriors. For example, students with high certainty will have smaller variance in their prior than students who have middling certainty.

Putting the pieces together. We now have each of the pieces that can be put together to generate an estimate of the distribution of both the likelihood and prior from a sequence of point estimates of the posterior. Diagrammatically, the process takes the posterior measure, applies inverse Bayes to generate a possible range of likelihoods, then uses the heuristic algorithm to narrow down the possible range of priors.

If this is done over a sequence of tasks, we can begin to build up a distribution that represents the student's prior distribution. It is then possible to generate an estimate of the posterior using this prior and likelihoods under different contexts:

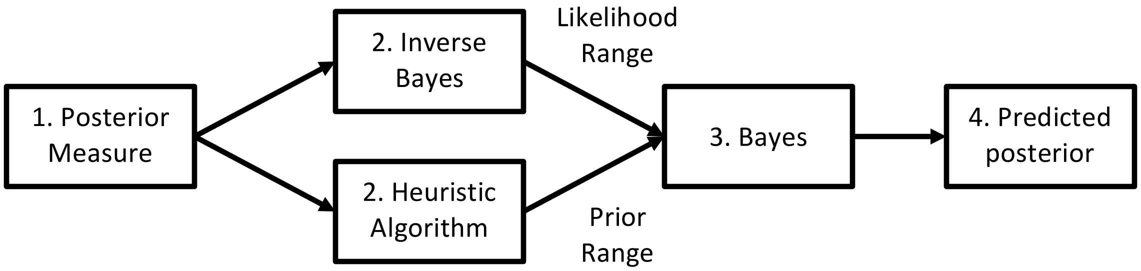


Figure 6. IBFi algorithm process.

Predictions can then be compared to future measures of the posterior to determine how accurate the model is as a forecasting method.

Appropriate Measures

The functionality of the process is dependent on finding meaningful measures for the posterior. The appropriate candidate would be something that is continuous and gives insight into the level of certainty a student has in their performance. Correct/Incorrect will therefore not suffice, but perhaps the cumulative correct/incorrect will. Alternatively there are several partial credit measures that exist and have found utility, such as Rasch models (Masters, 1982), EM Algorithm estimates (Muraki, 1992; Penfield, Myers, & Wolfe, 2008) and more recently the Electronic Tutor algorithm used by Wang and Heffernan (2010). The third possibility is the use of student certainty itself, such as used by Gardner-Medwin to judge proficiency on medical examinations (2013). This strategy essentially asks the equates the posterior with how confident a student is in her own answers.

Theoretical Consequences of Subjective Probability & Inverse Bayes

There are several theoretical consequences to both conceptualizing student knowledge as strength of belief and analyzing it using IBFi. In the first instance, it is unclear what constitutes guessing. The model must account for anything in the environment that might impact the student through the likelihood and their prior belief. For a student to guess in the sense that she is using no information there must be no environmental impact and their prior knowledge must be irrelevant. If this were a realistic proposition the student would then need to operationalize it through some form of random allocation of confidence to hypotheses. Given that people are notoriously bad at picking random numbers, such behavior is unlikely (Neuringer, 1986).

Alternatively the idea of uncertain belief intimates that *all* student performance is really guessing. Students are using their best estimate to make a decision without certainty. The consequence is then that guessing is likely not a useful construct to interpret student performance.

Knowledge as belief. The nature of the analysis is to measure strength of belief rather than correctness. A desirable result may well be a strong belief in a correct answer, but to limit analysis to this frame would restrict the range of inferences that could be made. Rather, it is whether the pattern of belief matches that of whoever designed the assessment. This is a fairly radical departure from assessment systems in which correctness is presented as objective truth, rather than a social construction. Imagining that students not only entertain misunderstandings, but also may need to be convinced to let go of those misunderstandings, is a substantial change to the way assessment is usually constructed (Moss, Girard, & Haniford, 2006). It is important however to view this not so much as stubbornness, although that is a possibility, but rather as a function of goal-directed utility. Students may not hold the same goals as the assessor, nor do they necessarily believe that the best hypothesis to apply to a task is the one that the teacher would choose. For example, you may get the correct answer by applying the knowledge that the most common answer is 'C'. There is utility in applying this hypothesis, especially if the aim of the student is to finish the task as soon as possible. It is not demonstrating the knowledge that the teacher or assessor is looking for though.

Ergodicity. The practical benefits of utilizing this method might well outweigh changes in the way that inferences are made from assessments. The prospect of using subjective probabilities for developing personalized inferences makes this method an attractive prospect in a world racing toward personalized learning through technology. In particular, it represents a source of variation that is not dependent on referencing a population of test takers; it is a non-ergodic measurement. The measurement issues mentioned earlier around personalization do not apply – subjective probability can make sense without reference to a student population. It holds value even without comparison to any other student.

Individualization. Because of its non-ergodic nature, inference made using the IBFi algorithm belongs entirely to the student who was measured, as estimates do not reference any other student. Such a property is of substantial use within personalization software as it allows much more fine-grained analysis of individuals.

Conclusion

In the middle of the last century several statisticians, philosophers and educationalists including de Finetti became interested in relating ideas about subjective probability to educational assessment. The effort never gained traction within education but did flourish in other fields such as economics and political science. It has been argued in this paper that subjective probability may yet find utility within education as means to automate individualization. Subjective probability may be useful for individualization efforts as an alternate source of within-student variation. This variation might then be leveraged to successfully differentiate student behaviors so that the most appropriate conditions are provided to each student based on a profile of how situational and aptitude complex factors interact for that individual.

Following this description the paper then outlines a methodological approach that utilizes subjective probability to estimate student aptitude and the contextual impact on the student. The key alterations to current methodology are described as relaxing the assumptions of Bayesian rationality and the adoption of a model based on the Inverse Bayes Formula. Details are provided concerning a) why relaxing these assumptions is practical and viable, b) how this process might be operationalized and c) and the kinds of utility it may have within the current assessment environment.

Personalization through the Application of Inverse Bayes to Student Modeling: Predicting Partial Credit with the Inverse Bayes Filter

Abstract

In this paper, I present a novel algorithm for predicting student performance within an Intelligent Tutor. The method is based on the Inverse Bayes Filter (IBFi), which predicts student performance utilizing within-student variation processed by the Inverse Bayes Formula. The goal of this paper is to determine the prediction accuracy of IBFi on real-world student data from an Intelligent Tutor. Two approaches are presented, 1) to look at patterns of prediction error when the algorithm predicts student cumulative average scores and 2) to compare the performance of the algorithm against the modified Bayesian Knowledge Tracing model (KTPC) of Wang and Heffernan (2011) when predicting partial credit scores. IBFi and its performance are demonstrated in the case of middle school students using an online math tutor, ASSISTments (*n* = 3684). Partial credit is calculated from student behavioral data within the tutor according to the method described in Wang and Heffernan (2013) and prediction accuracy is measured by root mean square error. The IBFi algorithm performs reasonably, out-predicting the KTPC model on a per-student basis but not across all predictions. The IBFi tends to over-predict high values of student performance and under-predict low values, and error is not distributed over skills equally. But the results demonstrate the feasibility of the idea of using Inverse Bayes to partition partial credit as an accurate way of predicting student performance.

Keywords: partial credit, Intelligent Tutor, Bayes, Inverse Bayes, prediction

Predicting Partial Credit with the Inverse Bayes Filter

Automated personalization of student interventions has been a long-standing goal of the educational enterprise. From Skinner's "teaching machines" to Khan Academy, the idea that technology can use individual student characteristics to ensure the most appropriate pedagogical approach is applied has inspired generations of educationalists and inventors. Computer mediated individualization, such as that found in Intelligent Tutoring Systems, requires that the process of measuring, validating and inferring student characteristics be automated to generate programmatic interventions. This paper investigates a novel algorithm, the Inverse Bayes Filter (IBFi), which utilizes individualized parameters to generate individual predictions of student performance. The ultimate purpose of this algorithm is that with reliable predictions of individual students come reliable parameters that can be used to inform automated judgments about personalization.

Introduction

Individualization

The process of automated personalization in Intelligent Tutors has largely relied on models that describe between-student item and skill level parameters rather than within-student parameters. The dominant Intelligent Tutoring model, Bayesian Knowledge Tracing (BKT) was originally conceived with an individual-level parameter but the implementation of this version has not been widely utilized (Yudelson, Koedinger, & Gordon, 2013, p.2). Individual-level parameters require the exploitation of within-student variance, which can be a difficult measure to acquire and interpret. In this paper I look at two sources of within-student variance: cumulative average and partial credit. The cumulative average is self-explanatory but partial credit is a more elaborate construction in need of some elaboration.

Partial Knowledge. The growing variety of task formats afforded by computer-based assessments is raising the possibility of utilizing a greater range of student behavioral data to draw more sophisticated inferences (Vonderwell & Boboc, 2013, p.24). Currently, successful student modeling approaches such as Bayesian Knowledge Tracing (Anderson, Corbett, Koedinger, & Pelletier, 1995) and Performance Factor Analysis (Pavlik et al., 2009) utilize student behavior in terms of a binary, correct/incorrect, input. But measuring student performance using a range of behavioral inputs may lead to useful inferences about learning and more effective interventions.

To expand the available behavioral inputs it is common practice to consolidate several measures into a partial knowledge scheme rather than grading based on correct/incorrect (Plano & Toby, 2004, p.180). In a partial knowledge scheme students are considered to possess incomplete information or understanding of a concept (Coombs, Milholland, & Womer, 1956). It stands in opposition to the dominant way that knowledge is modeled in educational assessment; as a binary “you either know something or you don’t”, referred to in the testing literature as complete, full or exhaustive knowledge (Falmagne, Cosyn, Doignon, & Thiéry, 2006, p.76). Whereas complete knowledge has an obvious scoring method to support it, Number Correct (NC) scoring, there is no definitive scoring method for partial knowledge, though there have been many suggested methods. Systems for scoring and making inferences from measures of partial knowledge have been investigated for over 50 years (Ben-Simon, Budescu, & Nevo, 1997, p. 65). Vygotskyian ideas of Proximal Development can be considered the grandfather of partial knowledge scoring methods, but there are many operationalizations in use today. Some use elimination of incorrect answers, "How many incorrect answers can the student identify" (Coombs et al., 1956), others use time, "How long did it take a student to perform a task?" (Bouffard-

Bouchard, 1990), or the level of assistance, "How many hints did the student need?" (Razzaq & Heffernan, 2006).

The chief benefit of pursuing partial knowledge through these scoring methods is the increased measurement variance that they provide, in particular within-student variance. Theoretically at least, more information should lead to a more detailed and precise characterization of the student (discrimination) and therefore more appropriate personalization (de Finetti, 1965). For example, non-desired behaviors such as guessing or gaming can be more readily identified or eliminated (Lau, Lau, Hong, & Usop, 2011). Less clear are the benefits that partial knowledge has in the prediction of future behavior. Recently, Bradbard et al. (2004) have reported gains in the ability to predict a student's future performance using Coomb's partial knowledge scoring method (1956), though they caution a need for more and better implemented research. Prediction of future student behavior is especially important in adaptive systems such as Intelligent Tutors. Improvements in the ability to predict student behavior can improve the automated system responses required to personalize student instruction through these platforms.

Models

Model Tests

The aim of this paper is to test the accuracy of IBFi at predicting student future performance. As a comparison model I have chosen the Partial Credit Knowledge Tracing model (KTPC) of Wang and Heffernan (2010). The two models have important differences but are similar in that they can utilize a continuous measure (partial knowledge) to predict student future performance.

Bayesian Knowledge Tracing. Although there are no widely used models that currently generate only individualized parameters, KTPC generates skill level parameters utilizing student partial credit. KTPC is a variant of Bayesian Knowledge Tracing (BKT) a forty year old model that was substantially developed by Corbett and Anderson (Corbett & Anderson, 1994). BKT is a form of recursive Bayesian estimation where the prediction is being made of whether or not a student possesses and can activate a given skill based on incoming correct or incorrect answers to items. It relies on two essential assumptions: 1) that items can reasonably be considered correct or incorrect, and 2) that collections of items can be attributed to a single skill. The model has four parameters that must be solved for or supplied:

- P(L0), the initial probability that the student knows a particular skill
- P(G), the probability of guessing correctly, if the student doesn't know the skill
- P(S), the probability of making a slip, if student does know the skill
- P(T), the probability of learning the skill if the student does not know the skill

BKT & Partial Knowledge. Bayesian Knowledge Tracing has had mixed results when utilized to predict partial knowledge. Yet, although it has not conclusively shown improved performance relative to prediction of binary correct/incorrect performance Wang and Heffernan have devised a modified BKT algorithm and scoring technique that reliably outperforms prediction of partial knowledge over the binary alternative: KTPC (Wang & Heffernan, 2010). This model is the same as the classic Knowledge Tracing model, but where the student performance node is a continuous partial knowledge score rather than a binary correct/incorrect score. All other parameters (guess, slip, initial state) are modeled as would be done in a conventional BKT analysis.

Inverse Bayesian Filter. A model type that is routinely used with partial knowledge to both define constructs and measure their change is the family of Rational Models. Rational Models describe and explain human learning, judgment, and inference, often through the use of Bayes Theorem. They are used to determine whether people's behavior is rational (whether it conforms to some optimality constraint) but if their implementation is inverted (i.e. – run backwards), instead of describing how a task *should* be completed, they generate the parameters that describe the way that people did complete a task (whether rationally or not). This inversion is the basic idea behind the process of the Inverse Bayes Filter.

IBFi seeks to determine the relative contributions of context and aptitude to student performance. Aptitude in this framework is whatever cognitive, emotional and conative resources a student brings to a task. Contexts are the conditions of the task that impact a student's performance. For example, a student may be certain about her name, but within a high stress context she may not be able to report it. Likewise, she may be very uncertain about the laws of thermodynamics, but if we provide enough context cues she may be able to choose the correct answer from a selection.

IBFi determines how knowledge and context should be weighted for a student, given their answer and according to logical probability. Bayes Theorem posits that the conditional probability of a hypothesis (posterior) is proportional to the product of the probability of that hypothesis (prior) and the likelihood of the available data conditioned on the hypothesis (likelihood):

$$P(hypothesis|data) \propto P(hypothesis) \times P(data|hypothesis) \tag{1}$$

Rational Bayes models such as Decision Theory (Schlaifer & Raiffa, 1961) and those of Griffiths, Kemp, & Tenenbaum (2008) treat the posterior as observed human behavior, the prior as stored knowledge and the likelihood as how the environment impacts the application of that knowledge (EG – the impact of the context). Bayes Theorem then gives the relationship between knowledge and behavior according to context. The following graph (Figure 1) demonstrates this by showing how, according to Bayes theorem, as aptitude increases (dot-dashed line), context (solid line) must drop quickly (become very hostile) to reduce performance (dotted line).

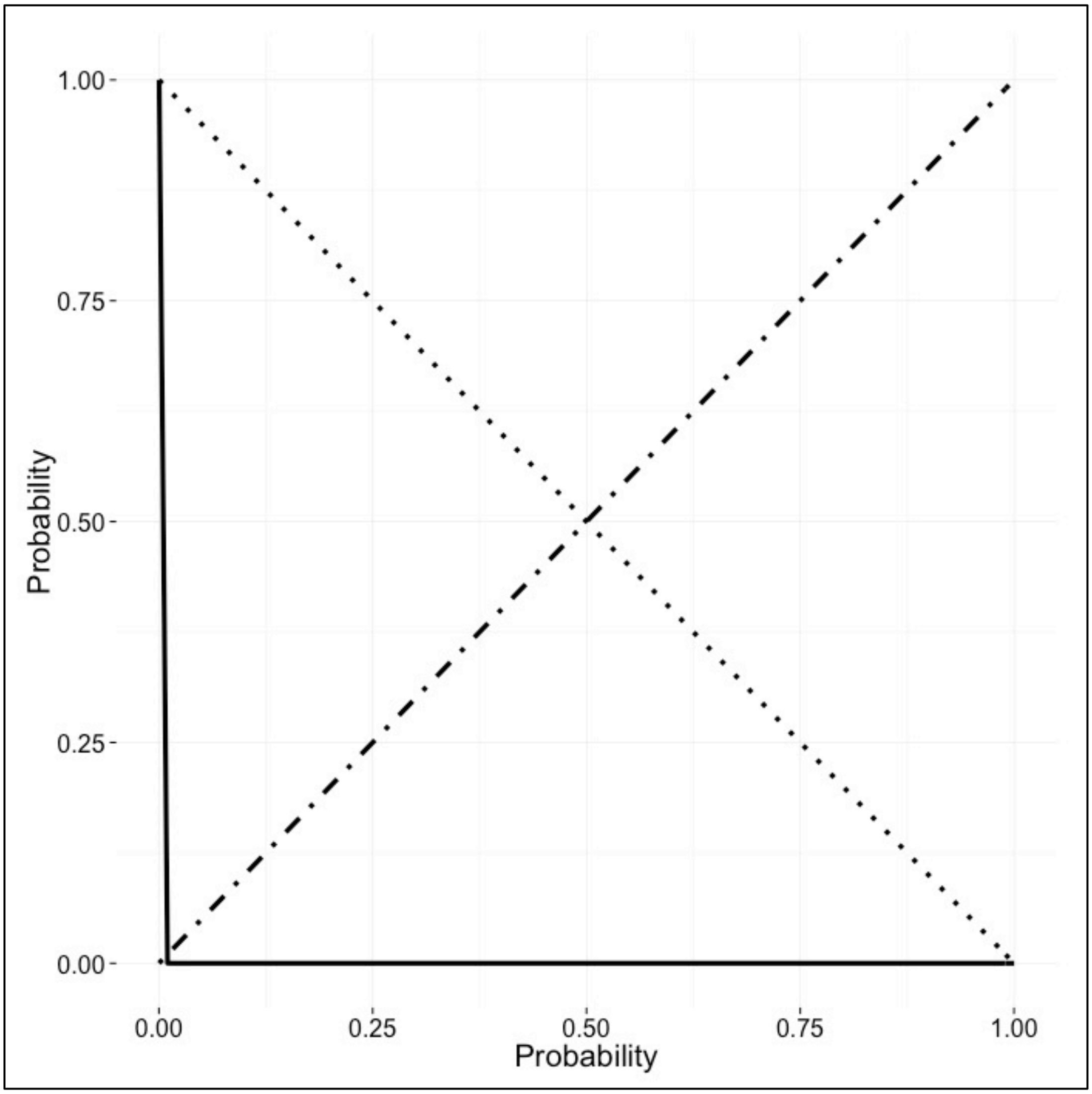


Figure 1. Theoretical values for posterior (dotted), prior (dot-dashed) and likelihood (solid) parameters demonstrating the balance between aptitude and context. For example, a student's partial credit can be high even if their knowledge is low, provided there are enough context clues. Likewise, partial credit can be low even if knowledge is high if the context is very unfavorable.

Rather than using this relationship to determine how a student *should* behave, IBFi use is to determine, based on a given posterior what the relative values of the prior and likelihood are. In other words, IBFi estimates the relative contribution to a student's performance of aptitude and context according to Bayesian rationality.

Methods

Comparing IBFi and KTPC

Both IBFi and KTPC are forecasting models that utilize Bayesian recursion to estimate parameters based on forecasting of a continuous input. For this reason it is fair to compare the accuracy of the two models, although they differ in important ways. KTPC utilizes between-student and between-skill variation to generate its parameters, while IBFi only uses variation within each individual over time. This difference has consequences for how we define the sample used to test the models. KTPC is *trained* on a subset of the data, during this training period the model sets its parameters based on the skill and student characteristics of students. It is then tested on a different, randomly selected subset of the data to measure its accuracy. This strategy cannot be employed to test IBFi, though, since it breaks the connection between students and their previous answers, the key source of variation used by the model to make predictions. Therefore a completely identical comparison strategy is not possible. Furthermore, since there are no models other than IBFi that exclusively use variation for individuals over time, it would not be possible to find a comparison model that did not have a similar problem. The strategy employed here to get as-fair-a-comparison-as-possible was to compare the two models on the same subset of students and to consider the IBFi training period as the number of sequential questions before the RMSE stopped changing in the third decimal place. In this way it is possible not to only compare the accuracy of the models but also how much training the algorithms require.

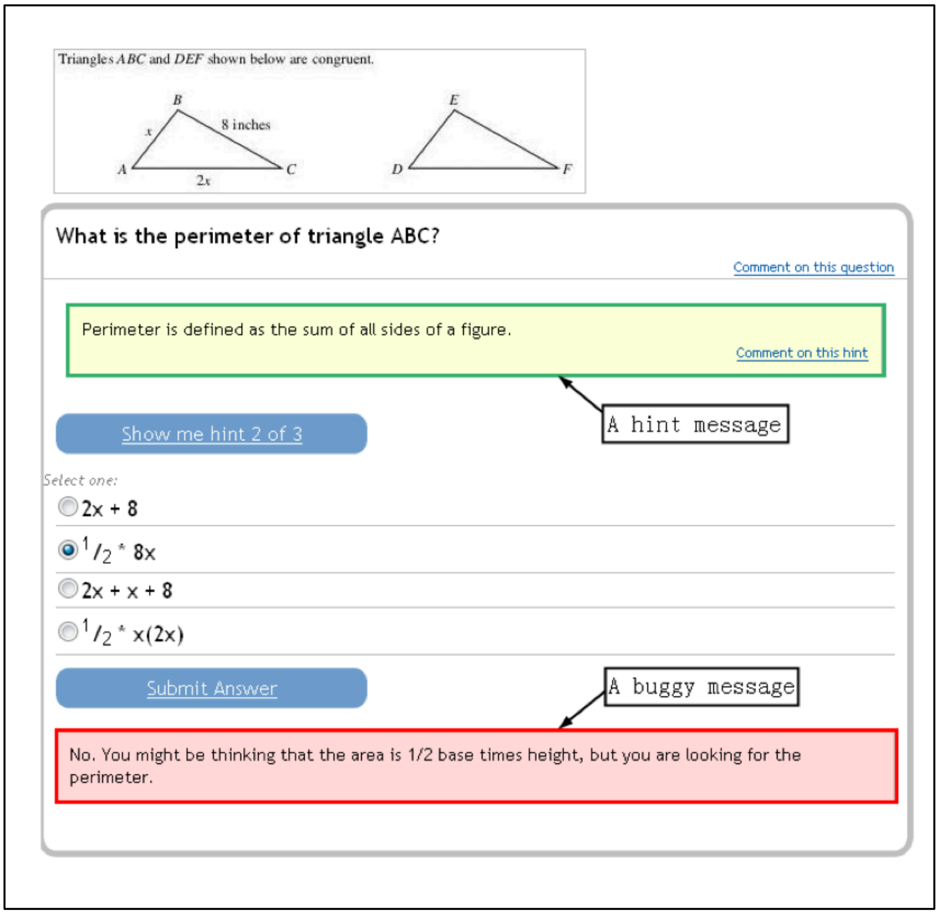
These models are also compared to a parameter-less model, simply using a student's previous score to predict their next score. Although there is no information gained from this strategy in terms of parameter estimates, and so it is of no use in the pursuit of individualization, this strategy provides a baseline level of accuracy that can informative.

Measures

Measures to feed into IBFi are a proxy for the combination of contextual and aptitude factors as construed as a probability value. They therefore need to be continuous, between 0 – 1 and reflect in some way a student's decision making at a given point in time. Two measures are tested in this paper: cumulative average score and partial credit. Cumulative average is the running average of correct and incorrect answers for a given student. The same partial-credit scoring regime that Wang & Heffernan demonstrated to test the reliability of KTPC is used here. This partial-credit scoring regime combines hints, scaffolding and correct answers to produce a partial credit score.

Data

The same data set with which Wang & Heffernan demonstrated the reliability of KTPC is used here. These data set consists of 3684, 12-14 year olds in the eighth grade of a school district in the North East of the United States. Student data was collected through ASSISTments, a web-based math tutoring system designed to prepare students for state standardized tests (Figure 2). Data consist of 211,342 item records including information about the number of hints and how much scaffolding (breaking a question into sub-questions) a student required to solve a given problem. No students can be identified and the research has been approved by CUHS.



Triangles *ABC* and *DEF* shown below are congruent.

What is the perimeter of triangle ABC?

Perimeter is defined as the sum of all sides of a figure.

Select one:

○ 2x + 8

● $^1/_2$ * 8x

○ 2x + x + 8

○ $^1/_2$ * x(2x)

No. You might be thinking that the area is 1/2 base times height, but you are looking for the perimeter.

Figure 2. Example task from the ASSISTments online tutoring system.

Results

The following results are presented in two sections that address the overall question, "Does the IBFi algorithm predict student performance?" The first section answers this question through analysis of the entire data set, examining patterns in error when forecasting student cumulative scores. Section two uses a subset of the data to investigate how the IBFi algorithm compares to another algorithm, KTPC, when forecasting partial credit scores.

1. Forecasting Cumulative Average Scores

Overall accuracy. Prediction of student cumulative average scores was analyzed through root mean squared error, and trends were identified across time and with respect to skill to determine whether prediction performance was associated with either variable. The overall accuracy for the 211,342 predictions as measured by RMSE was 0.1151 with a range of error in these predictions from -0.5-0.5 and the distribution centered close to zero (M = -0.023). A hex plot of error vs. predicted scores shows an unsurprising, slight linear relationship between high predicted values and over-prediction, and low predicted values and under-prediction, r(211340) = .037, p < .0001 (Figure 3). A *tail* can be observed at either end of the distribution in this plot. These tails represent the algorithm making large corrections in predictions over the first six questions from the seed prediction of 0.5.

Previous value prediction. One point of comparison is to see how the model compares to a prediction simply based on student's previous cumulative average. By using this method, a more accurate overall RMSE of 0.087 is achieved.

Time and skill. The dramatic improvement in prediction accuracy as the algorithm learns can be observed in a line plot of error vs. questions answered, identifying each student (Figure 4). Early predictions were generally poor, but improved as the algorithm processes more data. RMSE decreases from a high of 0.50 to a low 0.007, dropping from 0.50 to 0.12 within 10 questions. Error is not evenly distributed by the skill type required to answer the question though; some skills are associated with more error than others. The difference between the skill with the largest RMSE and the smallest is 0.212.

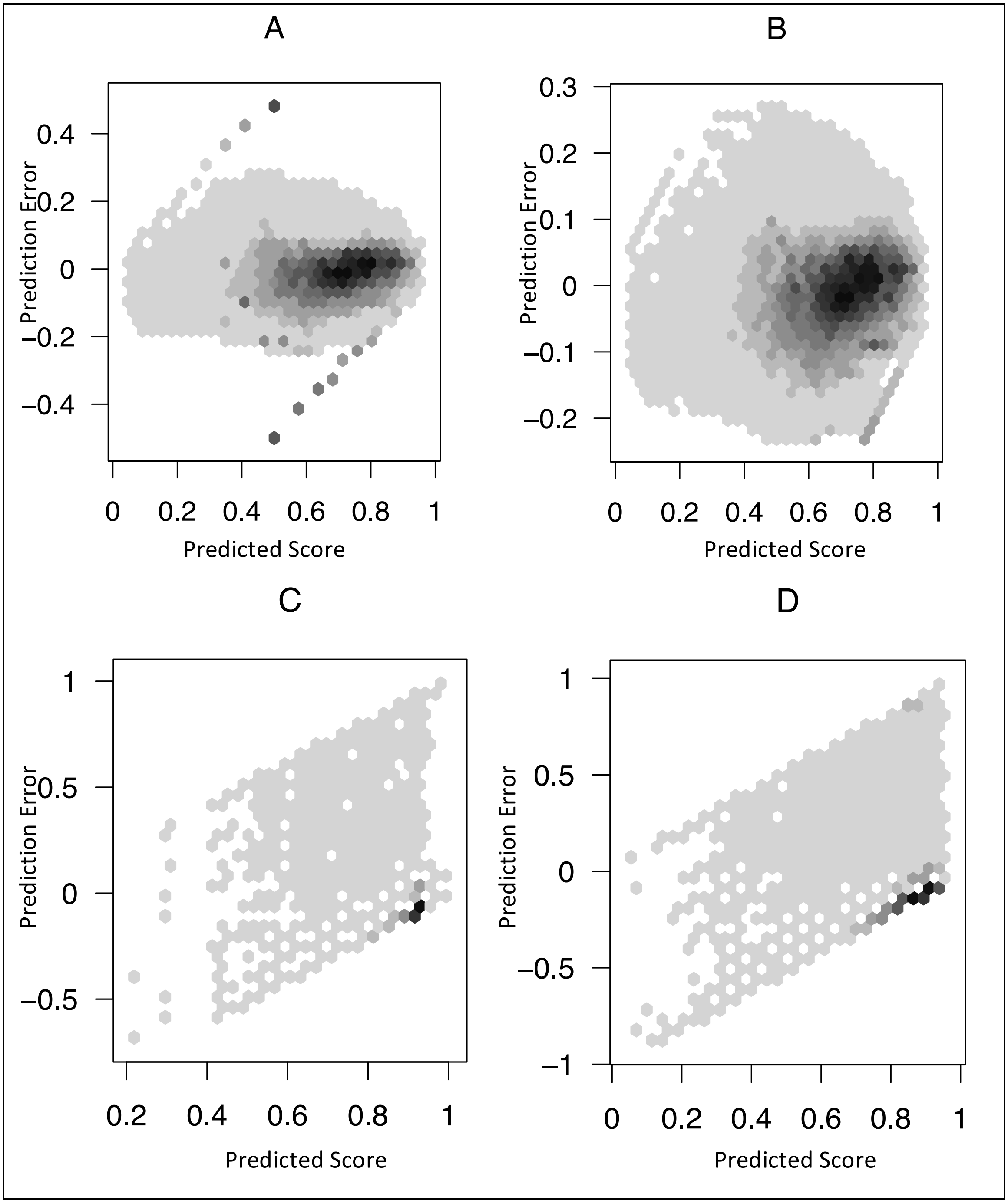


Figure 3. Hex Plots of prediction error vs. predicted scores. A. IBFi predicting cumulative average scores. B. IBFi predicting cumulative average scores after the first 6 predictions are removed to show that these questions represent the tails. C. KTPC predicting partial credit. D. IBFi predicting partial credit. Shading represents the number of predictions within hexagonal space.

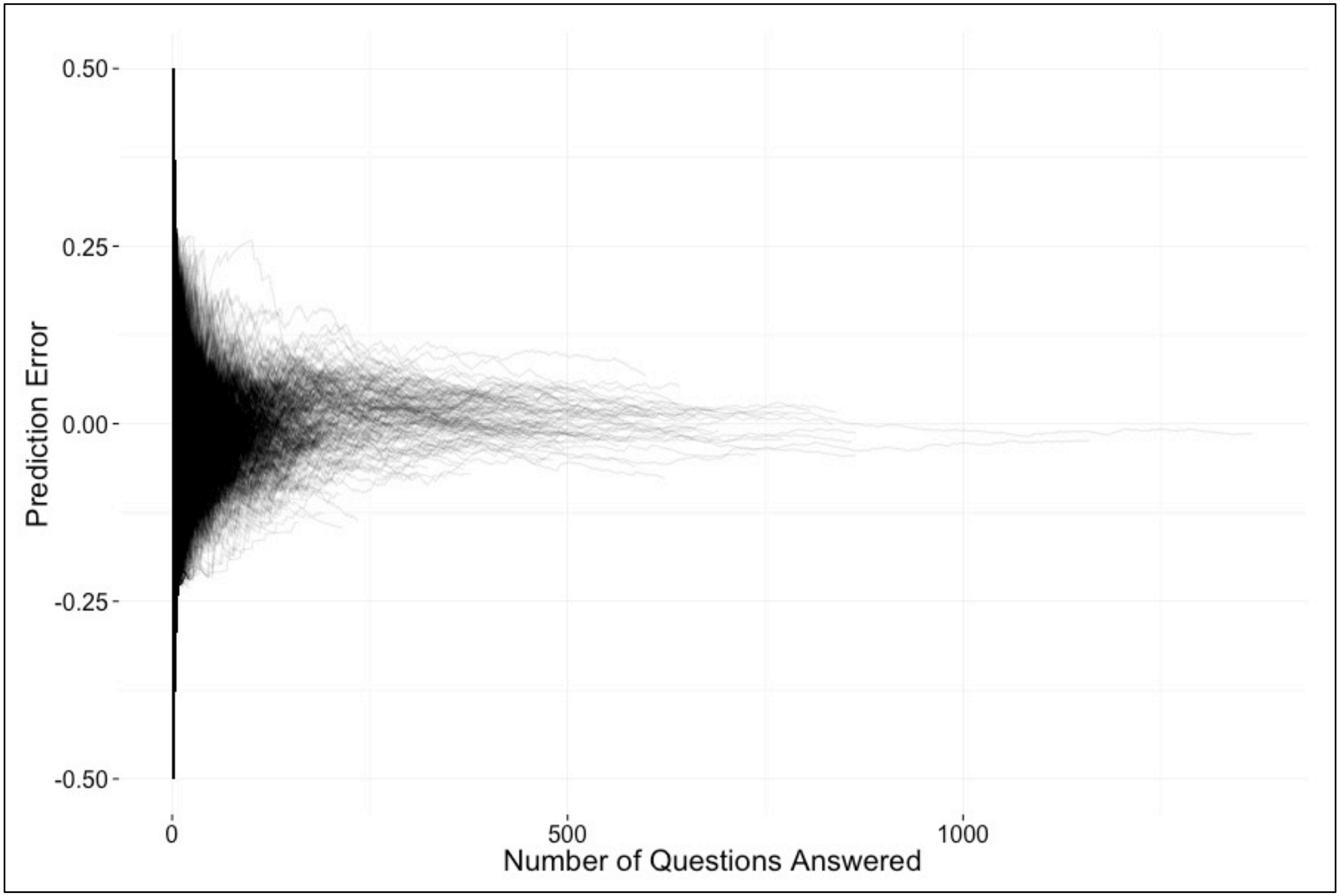


Figure 4. Line plot depicting prediction error vs. number of questions answered for the IBFi forecasting student cumulative average. Each line represents the error in prediction for one student.

2. Comparison between KTPC and IBFi

To further investigate the accuracy of the IBFi algorithm, it was compared to another continuous input, forecasting model, the KTPC algorithm of Wang and Heffernan (2010). The comparison was in terms of RMSE of predictions on a subset of the data for which Wang & Heffernan's partial credit measure had been calculated. Partial credit is more challenging to predict than cumulative average, as student scores can range from 0 – 1 from one question to the next. Accordingly, the variance of the partial credit scores is 0.04 compared to 0.03 for the cumulative average. The greater variance of partial credit makes it more difficult to predict simply by using the previous student score. The previous-value strategy generates an overall RMSE of 0.3598, compared to the overall RMSE of the IBFi algorithm of 0.3095 when tested on the entire data set?

Training Set. To determine the accuracy of the KTPC prediction model Wang and Heffernan *trained* their model on a random selection of 71% of the student data and then used it to predict the remaining 29%. This strategy is not possible with the IBFi algorithm as it is temporally dependent on each individual student – randomization at the level of the prediction will jeopardize this dependence and will change what the algorithm predicts. For the sake of comparison, though, the IBFi algorithm was applied to the same 2,313 students whose scores were predicted by the KTPC algorithm, and the time point at which the RMSE stabilized in the third decimal place was considered the *training* period for the IBFi algorithm, a period that consisted of the first 20 questions for each student.

Overall accuracy. The overall prediction characteristics of the two algorithms are similar, as can be seen in a hex plot of prediction accuracy vs. predicted values (Figure 3). Both models tend to have a similar error distribution, with over-prediction of high scores and under-prediction low scores and both have a preponderance of values for predictions that fall in the region around a partial credit score of 0.9, and a prediction error of negative 0.1. However, KTPC has a far larger number of predictions in that right hand region giving it a correlation of *r*(61786) = -0.043, *p* < .001, while IBFi has a correlation of *r*(61786) = 0.064, *p* <.001. The IBFi model also has a larger number of predictions in this region, pushing its average error values lower compared to the KTPC model (IBFi = -0.04, KTPC = -0.01).

Table 1

Root Mean Squared Error for Prediction of Student Partial Credit Scores on Training Subset

Model	Overall RMSE	Per Student RMSE	Per Skill RMSE	*n* predictions
KTPC[ł]	0.2824	0.3984***	0.2826*	61,789
IBFi	0.2922	0.2512	0.3100	150,186

*** Denotes significant difference between model RMSE where p < .0001
* Denotes a difference between model RMSE where p < .10, this is not considered statistically significant given the sample size of the data
ł Results and data generously provided by Yutao Wang, Worchester Polytechnic Institute

Model Comparison. The overall prediction accuracy using the training sample was slightly greater for the KTPC algorithm, and also slightly greater on a per-skill basis, *t*(150.8) = -1.835, *p* = 0.068 (Table 1), although IBFi was substantially more accurate on a per-student basis. A Welch two-sample t-test confirms that the difference between the student level RMSE distributions for the two models is significant, *t*(2227.933) = -35.152, *p* < .0001, with an effect size of 1.101. The IBFi algorithm also achieved this difference utilizing only 41% of the training data that the KTPC algorithm required.

Overall conclusions. The IBFi model performs reasonably well, though there is clear room for improvement with respect to overall error scores. (A tendency for the model to under-predict students with overall low scores and over-predict students with high scores.) Also some skills appear to be more accurately predicted than others. Despite these issues, IBFi outperforms a simple previous-score prediction when forecasting partial credit scores and it out performs the KTPC model on a per-student basis.

Discussion

In this paper, I present a novel algorithm for predicting student performance within an electronic tutor, IBFi, and quantify how successful it is at forecasting student scores from an Intelligent Tutoring system. Forecasting success is important if the parameters from the model are ever to be used to inform automated personalization strategies.

IBFi is based on Rational Models and utilizes the Inverse Bayes Formula to estimate the relative impact of student aptitude and contextual factors on student performance. Aptitude is comprised of the skills and information that a student enters a task with, while contextual factors are those things that influence her ability to demonstrate that knowledge. The calculations are made using partial credit scores and cumulative average scores, measures that do not assume binary correctness, but rather that a student can possess a spectrum of knowledge. IBFi treats the student as a Bayesian learner; her partial credit score is proportional to her prior knowledge and how she interprets her context. The Bayesian algorithm splits student performance into these two factors and then uses that information to make a prediction about the student's next score. This process generates estimates of a student's knowledge, the impact of context, and prediction accuracy.

The method and its performance are demonstrated in the case of middle school students using an online math tutor, ASSISTments. Partial credit is calculated from student behavioral data within the tutor according to the partial credit algorithm of Wang and Heffernan (2010) and then IBFi algorithm was used to sequentially predict those scores.

Key Results

In summary, there are two key results of interest and they seem somewhat contradictory:

1. Overall IBFi prediction accuracy is lower than for both the previous-score model predicting cumulative average, and KTPC predicting partial credit
2. Prediction accuracy was substantially higher on a per student basis for IBFi than for KTPC when predicting partial credit

IBFi and Previous-Score. With respect to the RMSE for overall predictions, there are several reasons why IBFi may be inferior to the previous-score method. IBFi effectively weights its predictions with past scores, so it is slower to adapt to changes in scores, and the starting point has a large influence on subsequent predictions. In contrast, using the previous score is a very fast adaption as the previous score becomes the next score. Such a fast strategy works well with cumulative scores, as they change quickly to begin with and then change becomes slower as each question has less impact on the overall score. However, with partial credit, this strategy is no longer as effective as the next score can vary substantially from the previous. Each subsequent partial credit score is based on more than simply the previous score.

KTPC and Skill. As with the previous-score method, KTPC has greater overall accuracy than IBFi, albeit with a much smaller margin (0.0098). There are many possible reasons for this but the two that were analyzed here (overall trend and trend by skill) suggest that it is a matter of accounting for skill rather than a better overall trend. The prediction trend for both algorithms is similar, with a tendency to over-predict high scores and under-predict low scores and a propensity of error at the lower end of high prediction values. With respect to skill, though, IBFi tends to have an uneven distribution of error across skill levels.

This distribution could be due to several confounding factors that are associated with the assignment of the ASSISTments program (sessions are assigned by teachers). However, it may also be related to IBFi having no explicit way of accounting for skills whereas KTPC does. KTPC weights predictions based on the skill that a question requires, IBFi only does this indirectly by picking up a signal that implies how proficient a student is at changing between skills in general, not any particular skill. It may be worth considering accounting for skill changes explicitly within IBFi as this may increase overall prediction accuracy.

The value of student-level parameters. Despite a lower overall RMSE, IBFi has substantially better accuracy on a per-student basis than KTPC. There may be several mechanisms within the two models that generate this difference, but one interpretation appeals to the variation utilized by each algorithm. KTPC uses between-student and between-skill variation across time to generate skill and student level parameters. IBFi only utilizes within-student variation across time. Predictions based on between-student (and subsequently between-skill) variation only include information about students relative to each other; they can lose information about patterns peculiar to individuals. However, IBFi may catch these differences with its within-student (or individualized) parameters and so generate more accurate predictions for individual students.

Greater accuracy of IBFi for per student but not aggregate scores begs the question, which is a *better* prediction, the prediction that does better at overall predictions, or the prediction that does better for each individual student. To some extent this depends on the aim of the prediction. If we are trying to predict the score of students in general or for a student for whom we know little of their history and we can reasonably assume they are similar to students in general, then KTPC will likely be more accurate. The ability to make a prediction of a student's behavior relative to other students is useful in this scenario.

However, if we have historical behavioral information for a particular student, and we wish to make a prediction for a single student (perhaps we only have data on one student), then IBFi is a better choice.

This second scenario seems to be the situation that arises in the case of personalization. For the purpose of personalization, having predictions that are tied as tightly as possible to the individual student is likely a useful property. IBFi is a first step toward developing models that may bring the benefit of individualized parameters to the personalization effort. The results presented here demonstrate the feasibility of the idea of using Inverse Bayes to partition partial credit as an accurate prediction algorithm that may further efforts to automate personalized education.

Personalization through the Application of Inverse Bayes to Student Modeling: Incorporating Student Certainty into IBFi Predictions of Student Performance in an Intelligent Tutor

Abstract

In this paper I validate the theory behind the Inverse Bayes Filter (IBFi) through testing whether student certainty (or confidence) improves prediction performance. The Inverse Bayes Filter utilizes the Inverse Bayes Formula to estimate the relative impact of student knowledge factors and contextual factors on student performance. This strategy is based on the assumption that the way students represent the world can be modeled as a probability distribution, and that as students answer questions they draw from this distribution. Several proxies for this probability distribution have previously been trialed, cumulative average score and partial credit. In this article a third approximation of the probability distribution is tested, student's level of certainty or confidence. The model's failure to process certainty successfully would undermine the assumption that student decision-making can be modeled in this fashion.

The inclusion of student certainty is shown to improve the predictive performance of the model relative to models that do not use certainty. By weighting partial credit scores with student certainty the per student root mean square error of prediction is reduced by 30%. This evidence supports the IBFi model and its underlying theory, indicating that students' judgments about their levels of certainty in their answers contains information that can be successfully identified by the model. The results further demonstrate the feasibility of the idea of using the Inverse Bayes Filter to process partial credit and predict student performance.

Keywords: Inverse Bayes Formula, student confidence, certainty, cognitive tutor, prediction

Incorporating Student Certainty into IBFi Predictions of Student Performance in an Intelligent Tutor

Introduction

Approximately every ten years for the last century researchers have advocated the use of probabilistic or certainty based educational assessment.[1] Such assessment strategies require students to provide not only the answer to a question, but also an account of how confident they are in their answer. Yet, despite periodic enthusiasm, this method only gained widespread adoption recently when the cost of implementation dropped due to improvements in software and mobile computing (Eser et al., 2012, p. 37).

Initial interest in certainty-based questions arose out of an interest in the mathematical formalization of subjective probability at the end of the 19th century (Estes, 1976, p. 37). Since 1913 researchers sought to apply these theories of judgment to educational assessments (Woodworth, 1915, p. 10). The motivation from the educationalists' perspective was to determine if querying student confidence could provide useful additional information about student performance (Echternacht, 1972). Over the last century the utility of confidence testing has been demonstrated in terms of test reliability (Ebel, 1965; Michael, 1968; Rippey, 1968), identifying guessing (Taylor & Gardner, 1999), separating students based on their level of understanding (Gardner-Medwin, 1995), increasing student understanding (Echternacht, 1972; A Gardner-Medwin & Gahan, 2003; Ramsey et al., 1987) and explaining answer changing (Skinner, 1983).

Despite continued interest, confidence based testing has not made the leap into mainstream pedagogy or measurement practices. This reluctance may in part be due to the criteria set down by de Finetti (1965) for confidence to be considered a legitimate

[1] For example, Coombs, Milholland, & Womer, 1956; Jr, Albert, & Massengill, 1966; Lichtenstein & Fischhoff, 1977, 1977; Ramsey, Ramsey, & Barnes, 1987; Rippey, 1968; Sia, Treagust, & Chandrasegaran, 2012; Skinner, 1983.

psychological measurement technique: that a scoring system could be devised that the student could not *game* to her or his advantage, and that the measurement unambiguously improved reliability and validity. With respect to the first challenge, although many systems were developed, none have become preeminent, and with respect to the second, the founders of modern psychometrics, Lord and Novick, dismissed any improvement as unlikely (Echternacht, 1972; Lord & Novick, 1968, Chapter 16).

However, interest in the theoretical nature of student confidence has still been pursued in psychological research in the form of self-efficacy. Self-efficacy is the "perceived capability to perform a target behavior" and makes up the key feature of Bandura's social-cognitive theory (1977). Self-efficacy has gained strength from its robust ability to predict many behaviors, including student enthusiasm and learning (Pajares & David, 1994) and is also compatible with Dweckian theories of student motivation (Dweck, 1986). Recently this combination of motivation and self-efficacy within schools has been pursued in electronic-tutor research to improve prediction algorithms (Mcquiggan, Mott, & Lester, 2008), indicate critical thinking (Bruning, Zygielbaum, & Grandgenett, 2001), and self-regulated learning (Azevedo, Johnson, Chauncey, & Burkett, 2010). Overall, the consensus of this research appears to be that understanding how students think about their own performance is beneficial to understanding and predicting their behavior more generally.

It is important to note the difference between certainty and confidence here; confidence is strength of belief that can be general, "She is a confident person", or specific, "She has confidence in her ability". Certainty, on the other hand, is the attribution of numbers to this strength of belief, "She is 50% certain that she will receive a letter". Since this experiment involves numerical attribution it will use the term certainty rather than confidence to describe students' strength of belief.

Research Goals

In the following article I seek to build on work both from the measurement literature concerning confidence and research into self-efficacy. It looks at the reliability of using student certainty as a measure in combination with a novel algorithm, Inverse Bayes Filter (IBFi). IBFi separates the factors that contribute to student performance into two categories, student aptitude and situational factors, and generates a parameter for each. It then uses those parameters to predict future student performance. The overall aim of this article is to test IBFi using student certainty data to determine if this improves IBFi prediction accuracy.

The importance of certainty. Whether or not certainty improves IBFi accuracy is important since IBFi is based on a theory that models student actions as generated from their own *subjective probability*. This subjective probability motivates their understanding of the world and therefore what they do. For example, when answering questions about math, students hold a subjective probability about their answer that will inform whether they attempt to answer or not. If their subjective probability is too low they may give up, or if it is too high they may not think through the problem properly. In short, subjective probability represents the summary of the processes that drives decision-making.

The Inverse Bayes Filter uses this basic idea to process student actions, if all actions are the result of an underlying probabilistic process we should be able to infer that process from those actions. What's more we should then be able to use this probabilistic description to predict future actions. IBFi does this by using sequential actions to estimate a student's subjective probability distribution with respect to a particular task. It does this by differentiating between that part of the sequential variation that belongs to the context from that which belongs to the student (their subjective probability).

The model process depends on having a proxy measurement for the student's subjective probability. This proxy might be any continuous, sequentially collected measure that is connected to the decision-making process. Although cumulative average score and partial credit measures have been successfully examined previously, another obvious choice would seem to be student certainty. Student certainty or confidence is a direct estimate of a student's subjective probability – it is her understanding of why she is making a particular decision. It is therefore important that IBFi is not only able to predict student certainty, but that the student certainty improves prediction accuracy relative to measures that do not use student certainty. If certainty does not improve a prediction then it is an indicator that the model does not operate as theorized, it may work for another reason, but it isn't the reason that informed its construction.

Analytic Strategy

Overall Strategy

To determine the impact of certainty on the accuracy of the IBFi model the overall strategy of this paper is to compare IBFi predictions with and without including student certainty. Certainty represents the closest available approximation to a student's strength of belief or subjective probability. The theory informing the model proposes that it is this subjective probability that the model will be most accurate at predicting. If model performance is worse or the same, with and without certainty, then this constitutes evidence that the theory informing the model is incorrect. Conversely, if the model is more accurate when certainty is included then it may provide evidence that the theory underlying the model is viable.

Student Certainty

It is worth examining the nature of student certainty to provide context to the IBFi model tests. Beyond descriptive statistics, this examination of certainty is done in two ways. The first is to determine what relationships exist between certainty and other variables of interest, such as hints and attempts, and the second is to treat certainty as the student's own estimate of their likely performance. For example, if a student has a certainty of 0.5, she estimates that she will be correct on half the questions she answers. Using this measure we can then compare the student's accuracy at knowing their future performance with that of the IBFi algorithm.

Weighting

Simply comparing predictions of certainty with those of partial credit or cumulative average is not a reasonable strategy for this data set as students were asked fewer certainty questions than regular questions. This was done so that students were not over-burdened by extra questions and to maintain the familiarity they have with the online system. As such, confidence has far fewer data points per student, which may mean that it also has less variation and is likely easier to predict. This ease of prediction would undermine any findings since increased prediction accuracy would be due to the sparsity of the certainty data rather than any property of certainty itself (e.g. That certainty represents the subjective probability of the student). To solve this problem, cumulative average and partial credit scores were weighted with student certainty to see if this impacted how well the algorithm could predict their values. In this way a comparison could be made between certainty and non-certainty conditions while maintaining the same number of data points in each group.

However, there is the possibility that weighting simply *smoothes* the data out, making it less variable and therefore easier to predict. To ensure that this was not the case, measures of variance within the weighted and un-weighted data were compared. Data were also weighted with random student certainty. In this way it is possible to determine if it is simply the weighting process that is responsible for any improvement in predictive accuracy. This weighting-dependent improvement might happen if having a limited number of weights (0.00, 0.25, 0.50, 0.75, 1.00) could reduce the amount of variability in the data making it easier to predict. For example, if a student oscillated between certainty of 0.75 and 0.25, and partial credit of 0.25 and 0.75 her weighted score would always 0.1875. This would be a much easier run of scores to predict than 0.25, 0.75, 0.25.

Comparison. The comparison between weighted and unweighted student performance data is measured in terms of Root Mean Square Error (RMSE) and Students t-test is used to compare the differences between RMSE distributions. Effect sizes, measuring the impact of weighting on prediction accuracy, are measured as the standardized mean difference in RMSE between the weighted and unweighted prediction samples.

Notes on the Inverse Bayesian Filter

The chief aim of this paper is to test the utility of an Inverse Bayesian processing algorithm to deal with student confidence data. The inverse Bayes algorithm seeks to determine the relative contributions of context and aptitude to student performance in accordance with Bayes Theorem. Knowledge in this framework is whatever cognitive, emotional and conative resources a student brings to a task. Contexts are the conditions of the task that impact a student's performance. For example, a student may be certain about their name, but within a high stress context she may not be able to report it. Likewise, she may be very uncertain about the laws of thermodynamics, but if we provide enough context cues she may be able to choose the correct answer from a selection.

Inverse Bayes determines how knowledge and context should be weighted for a student, given their answer according to logical probability. Bayes Theorem posits that the conditional probability of a hypothesis (posterior) is proportional to the product of the probability of that hypothesis (prior) and the likelihood of the available data conditioned on the hypothesis (likelihood):

$$P(hypothesis|data) \propto P(hypothesis)P \times (data|hypothesis) \tag{1}$$

Cognitive Bayes models such as Decision Theory (Schlaifer & Raiffa, 1961) and those of Griffiths, Kemp, & Tenenbaum (2008) treat the posterior as observed human behavior, the prior as stored knowledge and the likelihood as how the environment impacts the application of that knowledge (EG – the impact of the context). Bayes Theorem then gives the relationship between knowledge and behavior according to context. The following graph demonstrates this by showing how, according to Bayes theorem, as knowledge increases (green), context (red) must drop quickly (become very hostile) to reduce performance (blue).

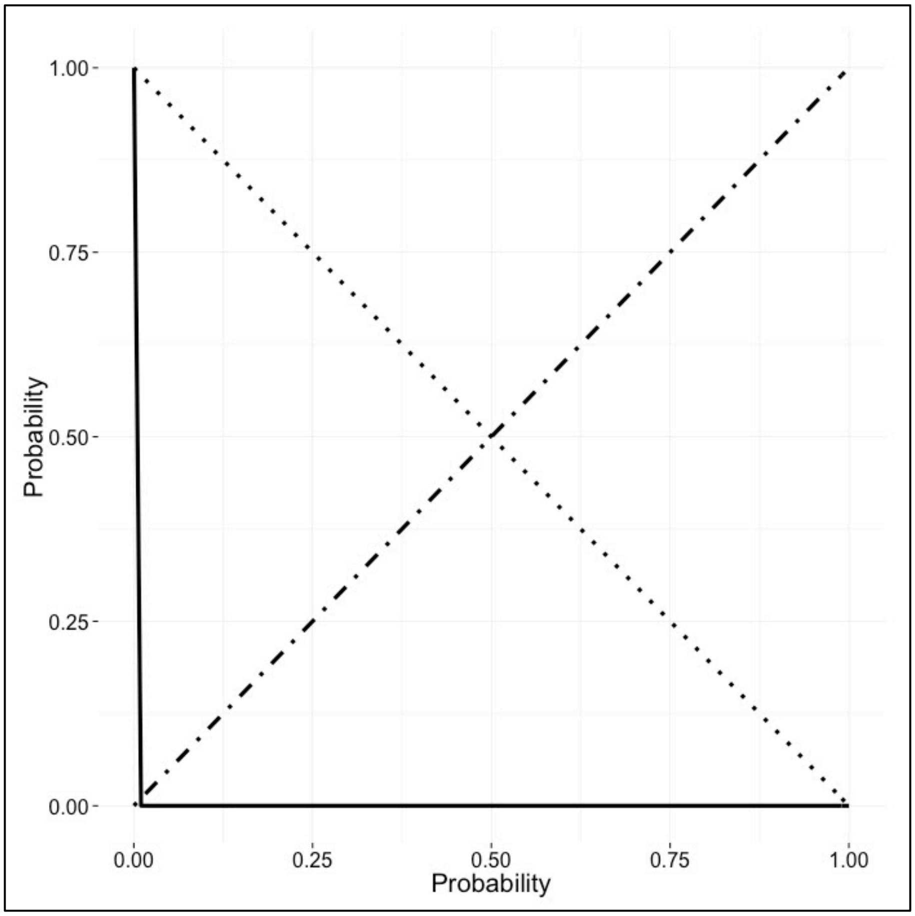


Figure 1. Theoretical probability values for performance (*p(h|d)*), internal aptitude complex (*p(h)*) and context (*p(d|h)*). Both *x*- and *y*-axes represent probability. We can interpret these values in terms of partial credit and knowledge. Partial credit can be high even if knowledge is low provided there are enough context clues, likewise, partial credit can be low even if knowledge is high if the context is very unfavorable.

An inverse Bayesian approach to this problem differs from these previous examples in that it does not treat the posterior as a stand in for a stimulus. For example, if an experimenter randomly assigned one group of students to tablet computers a Bayesian approach would treat tablet computers as a stimulus that had been equally applied to each student and differences in their performance would on average reflect the impact of the device. Differences between students in this model are treated as error. Conversely, in an Inverse Bayes model, we try to characterize these differences within the posterior, and they are interpreted as how each individual student is experiencing their context. In other words, in a Bayesian model all students have the same posterior and in an Inverse Bayesian model the posterior is free to range for each student individually.

It is important to note that there is an infinite number of possible prior and likelihood combinations for each posterior. However, the range of the likelihood will differ dependent on the posterior but the range of the prior will always be between 0 and 1.

To differentiate the range of the prior for each possible posterior, it is necessary to adjust the formula to account for how the variance of the prior might change dependent on its value. One way of achieving this change in variance is to assume that certainty has lower variance at extreme values. When students are very uncertain then they know they are uncertain; however if they are somewhat uncertain they could be certain or uncertain. We can do this by describing the variation in certainty as a Gaussian where the variance of the distribution is a quadratic based on the value of the mean of the Gaussian (Fig.4):

$$f(x,\mu,\sigma) = \frac{1}{-(\mu^2+(\mu-\left(\frac{\mu^2}{2}\right))\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2(-(\mu^2+(\mu-\left(\frac{\mu^2}{2}\right))^2}} \qquad (2)$$

The mechanics of the Inverse Bayes calculation are comprised of three steps. The first is to generate a seed prediction using mid-range values for the prior (range of 0-1) and likelihood (range of 0.33-1). This process generates the first *layer* in the weighted distribution that grows as we add more data. The second step is to add the difference between the prediction and the next observed student certainty measure to the prior. This new prior becomes the mean of a new range that will dictate the range of the likelihood, the two ranges will be added to their respective distributions. The third step is that the mean of these distributions is used to calculate a new prediction and the process starts again.

Example

The following is an illustrative example that brings together the ideas of certainty and IBFi within the experimental design. Consider a student who has the following partial credit scores after answering six questions:

0.1, 0.25, 0.4, 0.5, 0.8, 1.0

She has also given information about her certainty with respect to this type of question before her first and fourth question:

0.25, 0.50

We can therefore generate weighted and unweighted scores for this student:

	Q1	Q2	Q3	Q4	Q5	Q6
Unweighted	0.1	0.25	0.4	0.5	0.8	1.0
Weighted	0.1 x 0.25	0.25 x 0.25	0.4 x 0.25	0.5 x 0.50	0.8 x 0.50	1.0 x 0.5

The IBFi model can then be used to sequentially predict both the weighted and unweighted scores by using the previous score, processed through inverse Bayes. This processing would generate the following predictions and error values:

	Q1	Q2	Q3	Q4	Q5	Q6
Unweighted	0.1	0.25	0.4	0.5	0.8	1.0
IBFi Prediction	0.5	0.23	0.35	0.40	0.56	0.75
Error	0.4	-0.02	-0.05	-0.10	-0.24	-0.25
Weighted	0.025	0.0625	0.1	0.25	0.4	0.5
IBFi Prediction	0.5	0.04	0.044	0.09	0.10	0.34
Error	0.475	-0.0225	-0.056	-0.16	-0.30	-0.16

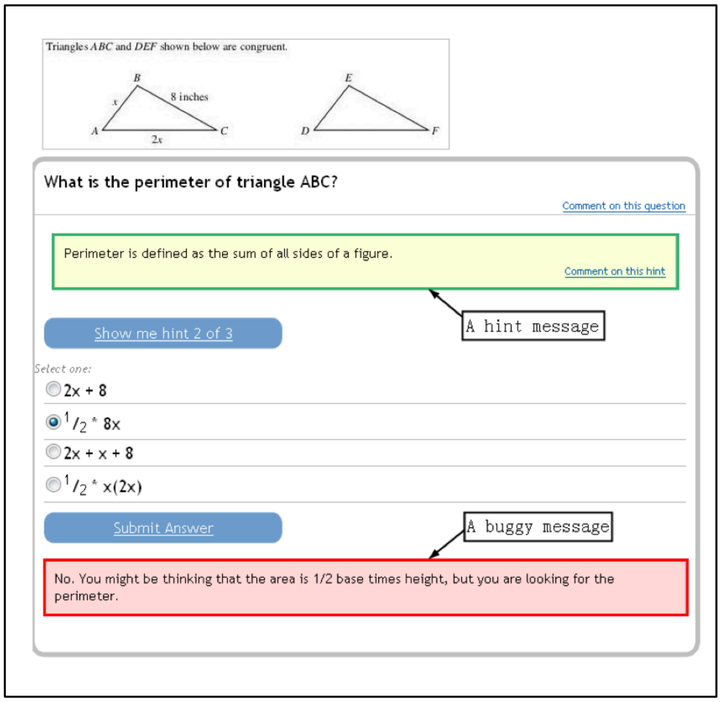
In this example, the overall error of prediction across the six questions for this student, as measured by RMSE, is greater for the weighted (0.25) than the unweighted (0.22) scores. In other words, IBFi has been less accurate at predicting the weighted than the unweighted scores. Such a result would imply that certainty weighting is more difficult for the algorithm to predict and that it is therefore unlikely that the theory underlying the model is correct, since that theory suggests that a closer approximation to a student's subjective probability will yield a more accurate result.

The remainder of the paper describes the results of repeating this example with 847 students across many questions.

Data & Measures

Data

The data set used for analysis consists of 847, 12-14 year olds in the eighth grade of a school district in the North East of the United States. Student data were collected through ASSISTments, a web-based math tutoring system designed to prepare students for state standardized tests (Figure 2). Data consist of 9,785 log records. Each record is comprised of an item ID, student ID, the number of attempts that the student took to complete the item, the number of hints they used in answering the item, whether or not the item was a certainty question, the student's answer and the average percentage correct that the student has attained on any items they have ever answered. This *prior percentage correct* includes all problem sets the students answered in the past, not just the present problem set, and this can include several years worth of questions. No students can be identified and the research has been approved by CUHS.



Triangles *ABC* and *DEF* shown below are congruent.

What is the perimeter of triangle ABC?

Perimeter is defined as the sum of all sides of a figure.

Select one:

○ $2x + 8$

● $^{1}/_{2} * 8x$

○ $2x + x + 8$

○ $^{1}/_{2} * x(2x)$

No. You might be thinking that the area is 1/2 base times height, but you are looking for the perimeter.

Figure 2. Example task from the ASSISTments online tutoring system.

Experience from the Student's Perspective

Two problem sets were designed around the multiplication and division of fractions and mixed numbers, using a Mastery Learning based structure called a Skill Builder. Skill Builder problem sets are unique in that students are randomly dealt questions from a skill bank until they are able to answer three consecutive questions accurately, thus 'mastering' the assignment.

Both problem sets were designed with two conditions: an experimental condition in which students were asked to self-assess their confidence in solving similar problems, and a control condition in which students were asked filler questions to control for the effect of spaced assessment. Random assignment was performed by the ASSISTments tutor at the student level. Throughout the course of each assignment, students were asked up to three self-assessment or survey questions. At the start of each assignment, students who were randomly assigned to the experimental condition were introduced to the skill of self-assessment, shown a set of problems isomorphic to those in the problem set, and asked to gauge their confidence in solving the problems using a Likert scale ranging from 'I cannot solve these problems (0%)' to 'I can definitely solve these problems (100%)'. Students who were randomly assigned to the control condition were polled on their current browser in an attempt to 'improve the ASSISTments tutor.' Only students who were in the experimental group were used in the current study.

Following these initial questions, students were given three mathematics questions. If students solved each of these three questions accurately, the assignment was considered complete. However, if students answered at least one of the problems incorrectly or they wished to continue answering questions, they would reach another self-assessment or survey question before being given another set of three math questions to try to master the problem set. This pattern happened a third time for students who were struggling with the content, until finally removing the self-assessment or survey element and simply providing back to back math questions until the student could solve three consecutive problems. Based on this design, high performing students may be asked to gauge their confidence only a single time, while students struggling with the topic were asked to reassess their confidence up to two more times throughout the problem set. The confidence question was always formatted using the same Likert scale, while the 'ASSISTments' improvement surveys changed slightly, polling students on various elements of accessibility.

Measures

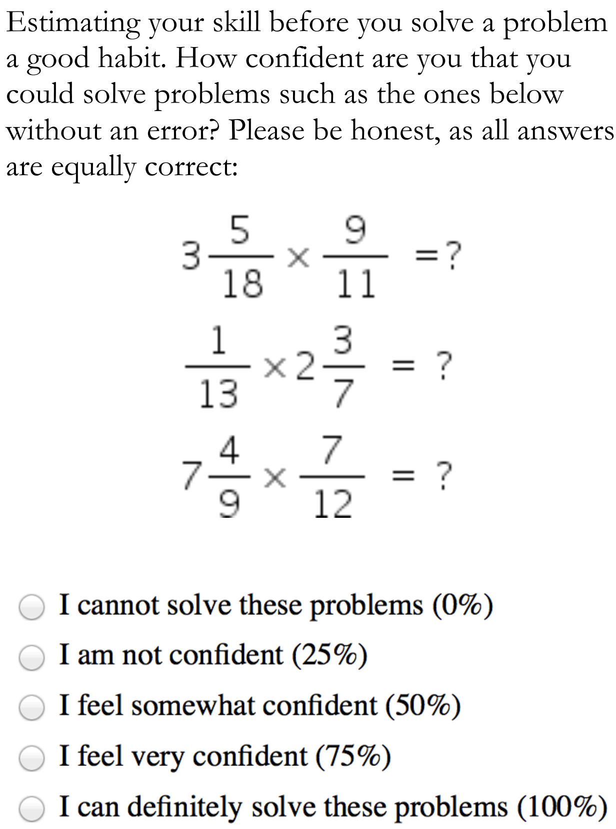
Student Certainty. Within the online tutoring system students were randomly assigned to an experimental condition in which they were asked about their confidence in the skill being tested every first, fifth and ninth item. Alternatively, in the control condition students were asked unrelated questions about the hardware they were using. Students were required to answer at least four questions to complete the problem set, though there is no limit to the number of questions that they can answer. However, after their ninth question there were no more subsequent certainty questions. A short instructional paragraph preceded the questions explaining how to answer the certainty type questions appropriately and the purpose of collecting this information. An example question is below:

Partial Credit. Several measures for student credit were utilized within this data set. Beyond whether a student got an item correct, information about the number of hints and how much scaffolding and how many times the student attempted a question was used. In addition to the inverse Bayes processing explained in detail above, this information was processed into a partial credit score using the same partial-credit scoring regime with which Wang & Heffernan (2011) demonstrated the reliability of KTPC.

Random weighting. It may be possible that weighting scores by a limited number of certainty values could reduce the variance of the data, making it easier to predict. To determine if this was the case, scores were also weighted with random student certainty values. Ten separate sets of certainty were generated by randomly drawing values from the original certainty data and then the mean results of these random draws are presented here.

Table 1 – Examples of treatment (confidence) style and control style questions.

Experimental Condition



Estimating your skill before you solve a problem is a good habit. How confident are you that you could solve problems such as the ones below without an error? Please be honest, as all answers are equally correct:

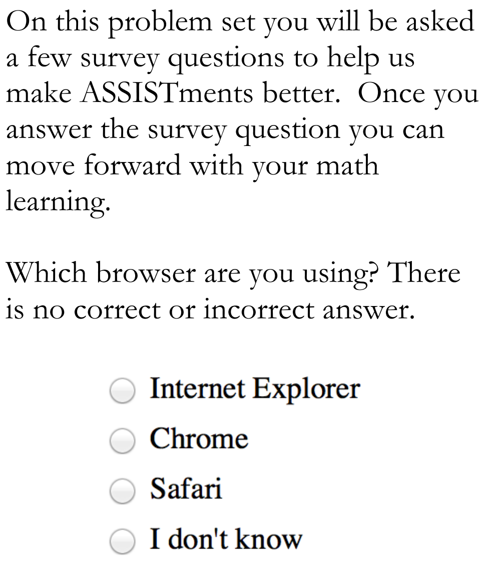
$$3\frac{5}{18} \times \frac{9}{11} = ?$$

$$\frac{1}{13} \times 2\frac{3}{7} = ?$$

$$7\frac{4}{9} \times \frac{7}{12} = ?$$

○ I cannot solve these problems (0%)
○ I am not confident (25%)
○ I feel somewhat confident (50%)
○ I feel very confident (75%)
○ I can definitely solve these problems (100%)

Control Condition



On this problem set you will be asked a few survey questions to help us make ASSISTments better. Once you answer the survey question you can move forward with your math learning.

Which browser are you using? There is no correct or incorrect answer.

○ Internet Explorer
○ Chrome
○ Safari
○ I don't know

Results

The following results are presented to address the overall question, "Is the IBFi performance improved by the inclusion of a student confidence/certainty measure?" Confirmation of the hypothesis that certainty data improves model performance provides evidence for the larger conclusion that the theoretical framework supporting the model is sound. The results are split into two sections, the first describes the certainty measure in detail, and the second compares model performance when the input is cumulative average score, partial credit, and certainty.

Student Certainty

The distribution of average student certainty is skewed to the left with a majority of students between 0.75 and 1.00 (*M* = 0.685, *SD* =0.251; Figure 3A). Average student confidence tends to drop between item 1 and item 9, though this is likely due to more proficient (and confident) students completing the three correct questions required to complete the problem set (M_1 = 0.7123, *SD* = 0.2640, M_5 = 0.6519, SD_5 = 0.2691, M_9 = 0.6264, SD_9 = 0.2818). The mean, per-student change in certainty decelerated slightly across items. Between item 1 and item 5 there was a mean, per-student change in certainty of 0.2376, while between item 5 and item 9 the mean, per-student change was 0.2163.

Other variables of interest. There are several variables of interest that are related to student certainty: hints, attempts, cumulative average score and the number of questions answered. The average number of hints a student used was 0.8155 (*SD* = 1.7333), the average number of attempts was 1.6333 (*SD* = 1.9649), the average maximum cumulative average score was 0.4836 (*SD* = 0.2816) and the average number of questions answered was 14.7533 (*SD* = 15.0734). These variables all had relationships with average student certainty, hints, $r(846) = -0.321, p < .0001$, attempts, $r(846) = -0.185, p < .0001$, maximum cumulative average score, r(846) = 0.3053, $p < .0001$ and the maximum number of questions answered $r(846) = -0.100, p < .05$. Though a plot of the relationship between average certainty and maximum number of questions answered suggests that highly certain students and very uncertain students answer fewer questions than mid- to high- certainty students (Figure 3B).

For some of the students the system has recorded the percentage of correct answers they have given over many problem sets in the past, sometimes stretching back several years. There is a relationship between a student's certainty at the beginning of the problem set and her prior percentage correct, $r(398) = 0.290, p < .0001$ (Figure 3C). There is a slightly less strong relationship between a student's certainty at the beginning of the problem set and their subsequent percentage correct, $r(846) = 0.245$, $p < .0001$ (Figure 3D).

Using student certainty to predict performance. As a comparison to the IBFi model, it is informative to compare its performance to student certainty. As a prediction of past performance, low certainty tends to under-predict performance and high certainty tends to over predict performance (i.e. prediction error is negative for low certainty values and positive for high certainty values: Figure 4A). The relationship between prediction error and certainty is described by a very strong correlation, $r(398) = 0.821$, $p < .0001$. A certainty of 0.75 appears to be the most accurate ($RMSE_{0.75} = 0.360$) and a certainty of 0.00 to be the most inaccurate ($RMSE_{0.00} = 0.487$) (Figure 4B).

As a prediction of future performance, the same pattern emerges. Low certainty tends to under-predict performance and high certainty tends to over predict performance (Figure 4B). The relationship is described by a weaker correlation than past performance, $r(846) = 0.453$, $p < .0001$. Though with future performance, students who were of mid-certainty ($RMSE_{0.5} = 0.356$) were more accurate on average than those who were very uncertain ($RMSE_{0.0} = 0.487$) or very certain ($RMSE_{1.0} = 0.485$).

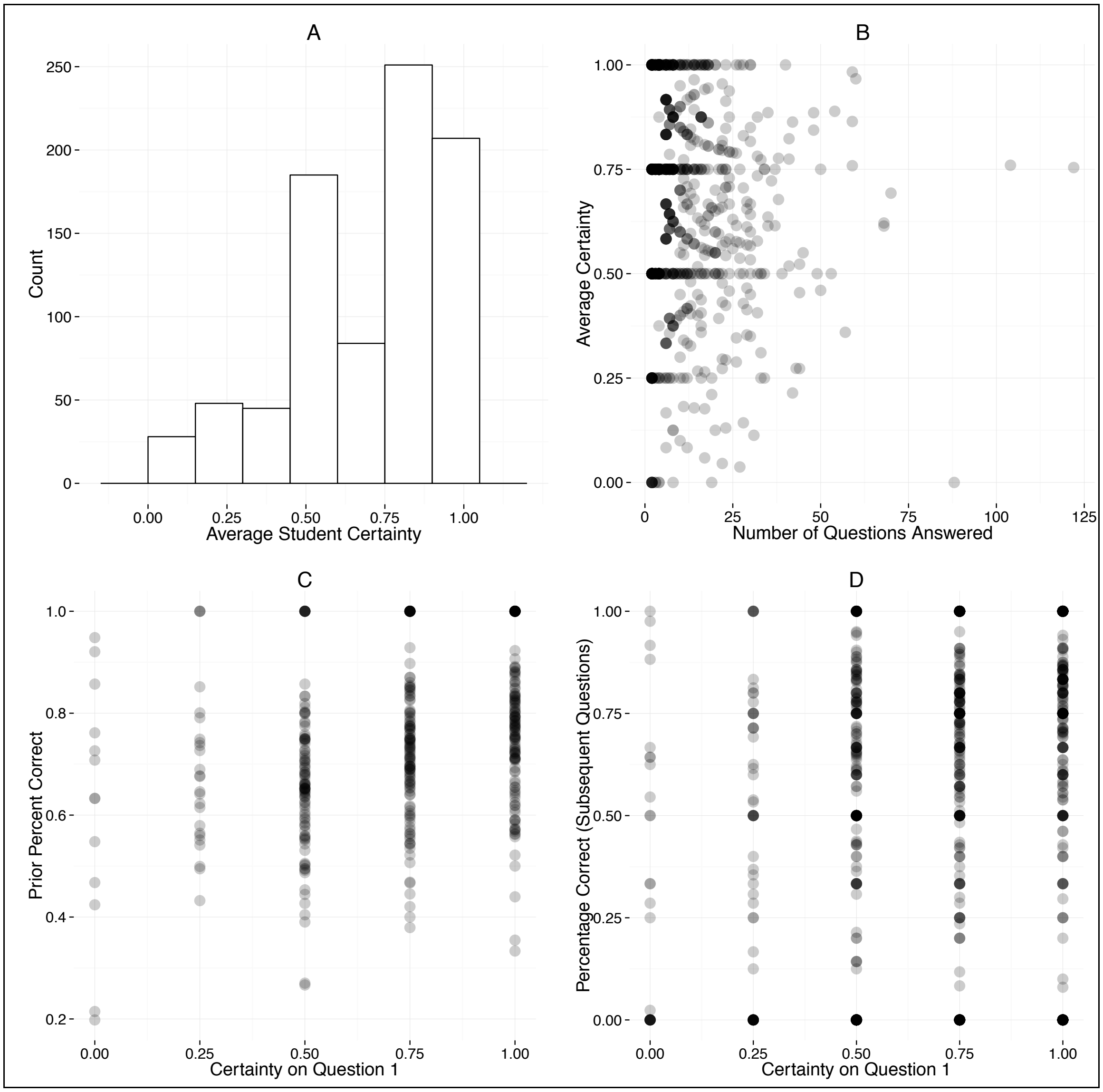


Figure 3. Characteristics of student certainty. A. The distribution of average student certainty. B. The relationship between average student certainty and the number of questions answered. C. The relationship between prior accuracy and student certainty at the beginning of the problem set. D. The relationship between accuracy during the problem set and student certainty at the beginning of the problem set.

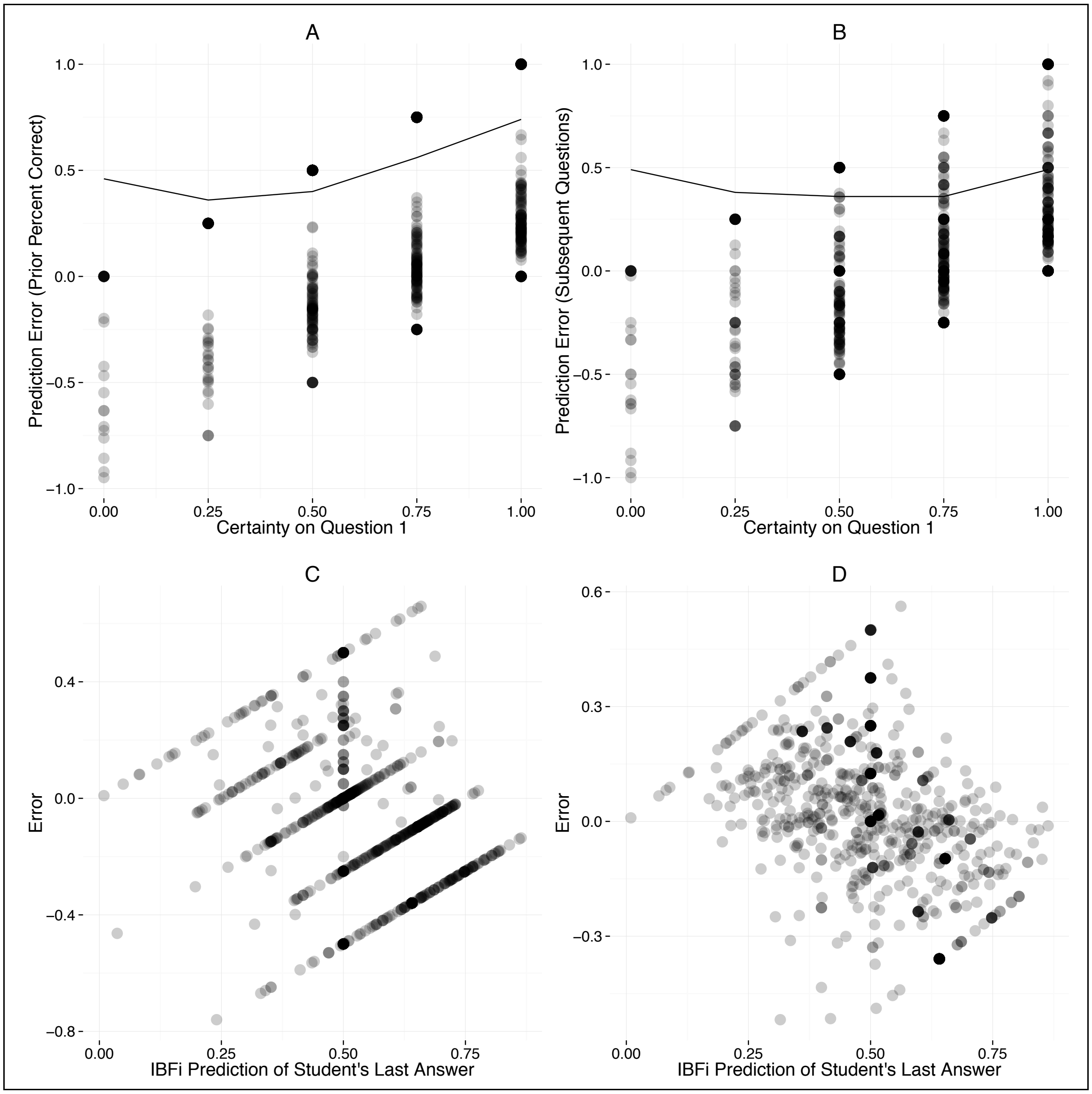


Figure 4. *Student certainty as a predictor of past and future performance compared to IBFi predictions of student final scores.* A. Past performance: prediction error of past performance plotted against student certainty at the beginning of the problem set. The line represents RMSE. B. Future performance: future accuracy plotted against student certainty at the beginning of the problem set. The line represents RMSE. C. IBFi prediction of final partial credit score and C. IBFi prediction of final cumulative average score.

Table 2. IBFi predictions of cumulative average and partial credit incorporating student certainty and random certainty.

	Overall RMSE	Per Student RMSE	Variance
Certainty Only	0.2992	0.2675	0.0737
Cumulative Average	0.1814	0.1869	0.0415
X Certainty	0.1779	0.1911	0.0677
X Random Certainty	0.2019	0.1892	0.0618
Partial Credit	0.3860	0.3974	0.1364
X Certainty	0.2832	0.2782***	0.1153
X Random Certainty	0.3755	0.3835	0.1218

**** Denotes significant difference between RMSE for certainty and non-certainty models where p < .0001*

The multiplication of certainty by both cumulative average and partial credit measures improved performance of the IBFi model with respect to student Root Mean Square Error (Table 2). The incorporation of random certainty did not improve the error rate to the same extent as the student certainty, though it did improve it on a per student basis for cumulative average. Weighting the partial credit scores with certainty does improve prediction error substantially, while weighting cumulative average has little impact at all. There was a marginal improvement in overall RMSE for cumulative average scores when weighted with certainty, and small decline in accuracy on a per student basis. For partial credit there was a substantial decrease in RMSE (29 times the improvement in accuracy compared to the improvement for cumulative average). There is also a statistically significant difference between per student RMSE for partial credit when weighted by certainty than without the weighting. We can think of this in terms of effect size, in that the strength of the effect of weighting partial credit scores on the accuracy of predictions using IBFi is 0.9231.

IBFi was also a far better predictor of final student cumulative average than the students were of themselves based on their initial certainty level. The RMSE for student predictions of their own cumulative average over the problem set was 0.4186, while the

RMSE for the predictions of IBFi for students' final cumulative average was 0.1937 (partial credit was 0.2530). Although students who had low confidence tended to under-predict their performance and students who had high confidence tended to over-predict their performance, IBFi, when using student certainty, had the reverse trend: low scores were over-predicted and high scores were under-predicted. This trend can be seen in Figure 4C and 4D. Figure 4C shows the prediction error when IBFi predicts the cumulative average of each student's final answer, and 4D shows the prediction error when IBFi predicts the final partial credit score of each student.

Discussion

The aim of this study is to determine whether student certainty (or confidence) can improve the prediction accuracy of the Inverse Bayes Filter (IBFi). This improvement is of interest since IBFi is based on the idea that students can be modeled as probabilistic reasoners – answering questions in the way that they believe is most likely to be correct. If inclusion of student certainty improves IBFi performance this goes some way to validating the underlying theory of the model; that certainty can be a direct window into the probabilistic reasoning of students. Characterizing student certainty gives some idea of how realistic this hypothesis is, pointing to certainty as a multi-dimensional summary of students' interaction with the online tasks.

Student Certainty

Certainty and decision-making. Strong negative relationships were seen between student certainty and other variables of interest such as hints and attempts; students who are more certain in their answers tend to require fewer attempts and hints to reach a problem conclusion. This intuitive finding suggests that certainty is related to the way that students navigate the online system in a fairly straight forward way; students who successfully complete problems are more certain, students who require more assistance are less certain. It is also possible that students of lower certainty are also more persistent than students with higher certainty though; high certainty students give up instead of seeking assistance. This alternate possibility is bolstered somewhat by the relationship between certainty and the maximum-number-of-questions answered by a student. Students who are very certain and very uncertain answer fewer questions than those with mid-high certainty. This trend also makes some sense, students with high certainty are also students who are goal oriented, and once they reach the goal they move on while students who are very uncertain are students who find the work difficult and give up quickly. Middle-high certainty students meanwhile are those that persist beyond the minimum goal requirement, perhaps to satisfy their own sense of understanding.

The finding that middle-high certainty students tend to persist while high-certainty students do not resonates with the work of Dweck (1986, 2012) on performance orientation. According to Dweck, students with a fixed mindset tend to be essentialists – they believe that their abilities are fixed traits that dictate performance. In contrast, students with a growth or incremental mindset believe that ability can be developed through learning and practice. Students who have a fixed mindset tend to work in a performance-oriented manner, where performance is a direct indicator of their general ability ("I was wrong, I am stupid") while students with an incremental mindset see performance as a reflection of an ongoing process ("I was wrong, I need to learn more"). The difference between the very high/very low and medium-high certainty students with respect to persistence may describe these different mindsets. Very high and very low certainty may be a proxy for a fixed mindset – a student who has a fixed mindset believes that she either has the ability required to perform or not, her certainty choices are limited to either zero or one. Within such a schema incorrect answers are a reaffirmation of a lack of ability, making persistence illogical. However, students with a growth mindset will rarely rate their certainty as 100% as they believe that there is always room to improve. Instead, they may rate their certainty as mid-high, but persist when they are wrong as they see it as a learning opportunity. The tendency of growth mindset students to persist while fixed mindset students do not is confirmed by Bandura, who suggests that growth mindset students are better able to deal with stressors than fixed mindset students (Bandura, 1993).

Relationships between student certainty and hints, attempts and persistence suggest that student certainty is related to the way that a student navigates the online system. Students are choosing certainty values in a consistent way that is connected to the tasks they are doing, not in a random way or in a way that is unrelated to the choices that are available to them such as would be expected if they were not taking the confidence questions seriously. This association supports the use of certainty as an appropriate measure for IBFi; there is information that can be extracted from certainty that pertains to the way students work through the computer based tasks.

Certainty and accuracy. There is a positive relationship between student certainty and both past and future performance. The relationship between certainty and accuracy suggests a stronger relationship with prior experience than with future performance. Students who are more certain at the beginning of the problem set are those who have been more accurate in the past, while those who were less accurate are less certain. There is a similar, but smaller relationship, with future performance; students who are more certain are more accurate during that problem set and students who are less certain are less accurate. However, this relationship is tempered by a substantial number of students across all certainty ranges who score as low as possible during the problem set. These students are over-predicting their future performance and they are not students who are simply exciting the program after starting the first question. It is worth looking at the error patterns of prediction then to see where these prediction errors are accumulating.

If we treat certainty as a prediction of past and future performance there is a clear trend toward low certainty students under-predicting their accuracy and high certainty students over-predicting their accuracy. This trend is somewhat explained simply by the possible amount of error a prediction can have – there is a greater possible distance between a high prediction and a low result than between a middle prediction and low result for example. If a student consistently predicts middling values then she will have, on average, less error if her observed scores are extremely high or extremely low. The tendency to reduce error when this strategy is employed may well explain the preponderance of students who sit at either end of the spectrum. It does not necessarily explain the spread of students across the entire range of error though, nor does it explain that mid-high students (certainty of approximately 0.75) being the best predictors of both past and future accuracy. If it were simply a matter of making the safest bet possible, students with predictions of 0.5 would be the most common and most accurate.

When we compare the correlation of past and future predictions with their error rates students have more conservative predictions of past performance and more optimistic predictions of immediate future performance. That is to say, the slope of the relationship is steeper and over a smaller range for past predictions (-0.98 – 0.54) than for predictions of future performance (-1.0 – 1.0). It appears that students rely on past accuracy to inform their certainty, which tends to be more optimistic about their future performance than is warranted *in this case*.

Trends in accuracy and error also seem to support the model. Students who are uncertain perform worse than those that perform well, and student performance is informed by past experience. IBFi works under the assumption that students model the world according to their certainty, and certainty should be weighted by past experience and current conditions. It also weights prior experience more strongly than current conditions. According to these results the stronger relative impact of prior knowledge seems to be the general strategy of the students. Prior experience informs certainty to a greater degree than current conditions, if it were reversed then we would expect to see equal error or greater accuracy for predictions on the current task than with past performance.

IBFi Accuracy

The more substantial test of the hypothesis that IBFi describes student behavior in terms of certainty is to utilize certainty as substrate for the algorithm. To test whether certainty made a difference to IBFi prediction accuracy, a comparison was made between predictions of student performance measures with and without inclusion of student certainty. Performance scores (cumulative average and partial credit) where first predicted unaltered and then predicted after they had been weighted by student certainty. The results were further compared to random weightings of scores to determine whether simply weighting the scores made a difference.

Improved predication accuracy (with caveats). The overall result from testing IBFi using certainty is that certainty does improve prediction accuracy with two important caveats: a) that improvement is substantial for partial credit and minimal or non-existent for cumulative average and b) consideration that the improvement is simply an artifact of the weighting process and has little to do with the certainty values themselves.

Cumulative average vs. partial Credit. With respect to the first issue, we need to explain why there is such a marked improvement in partial credit prediction with the use of certainty, and so little in the case of cumulative average. To some extent this is likely the result of information loss. Cumulative average is a very sparse measure with little variance. This sparsity makes it easier to predict (RMSE is far lower than for partial credit) but also that there is little variance with which to distinguish one prediction from the next. Although weighting cumulative average with certainty increases the variance somewhat, it isn't a large enough signal for the model to pick up a substantial gain in performance. Indeed, random weighting of cumulative average appears to have an equivalent effect.

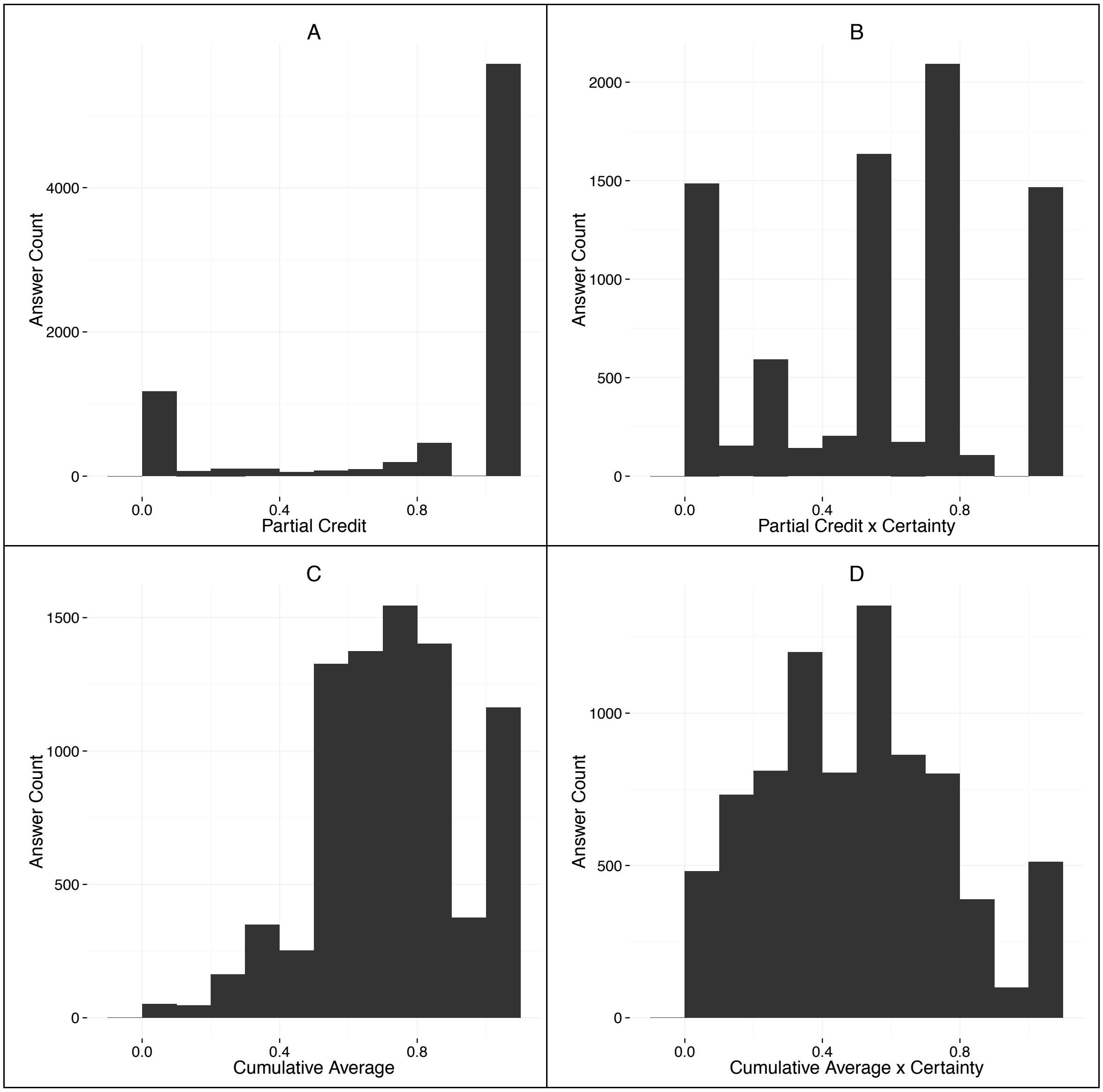


Figure 5. Histograms of weighted and unweighted partial credit and cumulative average scores demonstrating that that the weighting process should not make the scores easier to predict.

Conversely, partial credit is a richer measure, it is a multidimensional-summary of several aspects of the online task (accuracy, hints, attempts), and has double the variance of the cumulative average. The larger amount of variation distinguishes student behavior from one question to the next in more detail, producing a stronger signal for the algorithm to pick up. That certainty improves prediction also implies that it is offering information that is not already summarized by the partial credit measure. This extra information reflects support for the theory that informs the model, though we must account for the possibility that it is the result of a statistical artifact.

Accounting for certainty weighting. It is possible that prediction accuracy increases because the weighting procedure simply makes the values easier to predict (EG - it smoothes the values out). This smoothing is a legitimate concern, as we are looking for evidence that supports the underlying model, not only for a boost in performance.

The first piece of evidence that this is not simply an artifact is that the differences in variance between unweighted and weighted values do not differ greatly, and in the case of cumulative average, the variance on the weighted scores is greater than on the unweighted scores (Table 2). Indeed histograms comparing the distributions of weighted and unweighted scores confirm that the weighted scores, if anything should be more difficult to predict since, in the case of partial credit, the possible answers to predict shift from predominantly 1.0 to a range across 0.0 – 1.0 (Figure 4). It seems that a smoothing effect of weighting the scores is unlikely to have made prediction easier.

Comparing Prediction Error of IBFi to Students

Compared to students' predictions of their own success (as determined by their initial certainty), the Inverse Bayes Filter compares very favorably. IBFi has an RMSE of less than half that of the students. We might conclude that what the algorithm can learn about students over a few questions is far greater than what the students themselves have learned about their own behavior over all their experience, though this would be an over statement. If student certainty is not meaningful, then it would not aid the model in making predictions. Far more likely is that the meaning of the raw certainty value is contextually specific to the student and requires interpretation. IBFi provides such an interpretation, parsing it into situational and aptitude components. Taking student raw certainty as a prediction of future performance neglects the different meanings that the measure can take on – whether it is a prediction of the current context, or a prediction of past performance.

Conclusion

In conclusion, this paper sought to address the overall question, "Is Inverse Bayes Filter performance improved by the inclusion of a student confidence/certainty measure?" Confirmation of the hypothesis that certainty data improves model performance provides evidence for the larger conclusion that the theoretical framework supporting the model is sound. Since certainty is theorized to be a proxy for a student's subjective probability, if the model did not improve with the addition of certainty data then this would raise questions about the underlying theory – IBFi is not a working based on modeling subjective probability. However, since IBFi performance was improved by weighting scores with student certainty this suggests that the underlying theory is sound.

The results described here provide evidence that indeed student certainty data do improve the model prediction values substantially (more than 30% for partial credit). This improvement confirms the theory that the model describes individual students' subjective probability and that modeling students in this fashion can be effective. The results in this paper go a short distance down the road of validating this model. Models that are based on individual difference over time may prove useful in producing accurate personalization software that can adapt probabilistically to student behavior.

Personalization through the Application of Inverse Bayes to Student Modeling: Concluding Bookend

Concluding Bookend

This set of articles has described the theory and implementation of the Inverse Bayes Filter, a rational model that predicts student behavior based only on individual level variation over time. This novel approach to predicting performance is the only algorithmic approach with this strategy applied to educational data and this set of papers represents the first tests of its accuracy.

Findings

The substantial findings from the two empirical studies presented here are that the Inverse Bayes Filter:

- Can predict student behavior with increasing accuracy as more data are supplied by each student
- Can outperform the KTPC model *on a per student* basis, but not overall
- The model is more accurate at predicting partial credit than cumulative average score
- The model tends to under-predict low values and over-predict high values
- Weighting partial credit with student certainty (e.g. confidence) increases accuracy of predictions for partial credit but not cumulative average

With respect to the use of rational models, such as those based on Bayes Theorem, this is not an unequivocal endorsement, but neither is it a defeat. It represents some promise for the method and methods like it. The algorithm is close to the overall accuracy of a validated model, and is more accurate in a way that is predicted by theory – it is better at predicting the behavior of individuals. It also behaves as expected in that it is more effective using something that more closely resembles subjective probability, student certainty.

That is not to say that there are not serious questions that need to be pursued with respect to further validating this model, even beyond the obvious need for many more similar studies to demonstrate repeatability. The model does not perform as well as it might when predicting cumulative average score. This failure needs to be investigated thoroughly to determine whether it is an issue with the algorithm or whether the measure is unsuitable and why.

Further investigations are also needed to elaborate on the meaning of the parameters. Do these parameters actually represent internal (aptitude) and external (situational) factors that impact students? Do they help explain student behavior?

Beyond these basic questions about the parameters, there are ways that the model could be extended. It may be possible to further subdivide the parameters to hierarchically account for the structure of student behavior. For example, can we hierarchically organize the internal parameter with content parameters so that we can identify that topics such as fractions and decimals fit within rational numbers that fits with mathematics?

Further improvements may be possible with respect to performance of the algorithm through utilizing Markov Chains to make estimates and relying on distributions rather than point estimates of probability.

What does this mean for educators?

The ability to better predict individual level behavior is of consequence for educators in several different ways. Most obviously it is of consequence for efforts to automate individualization, the ability to better characterize individuals so that they can receive targeted interventions is at the heart of the personalization enterprise. IBFi type models could improve educational practice, from making better technological products that can scale more easily, to making more accurate inferences about students' needs. Yet this is only part of the personalization impact.

Rational approaches do not only open the possibility of better predictions but the possibility of different kinds of predictions. In the case of the Inverse Bayes Filter the use of a rational model allows the characterization of students based on their own past performance over and above their performance relative to their peers. This novel use of variation may well produce different kinds of categories to those that we currently use. These new categories may contradict current categories and understandings of the way students learn. It may allow different understandings of how interventions variably impact students that can then inform practice. With respect to IBFi in particular, the ability to parse student performance accurately into internal and external components may allow more informed decision-making around how classrooms are structured and how progress is defined.

Rational Models

Rational models like IBFi may well be the future of educational assessment, simply because of the explosion in the amount of educational data that is now available. As has happened in astronomy and the biological sciences, education and the social sciences more broadly need to revise the way that data are processed to deal with the variety, speed and amount of data that are now available. This increase in data means that sample based statistics, although they are still useful, need to be augmented by processing algorithms such as Inverse Bayes. As an example, Bob Williams at the Hubble Telescope moved from using sampling statistics to characterize the night sky, as had been done by astronomers for generations, to processing as much information as was possible with algorithms before statistical astronomers even saw the data. To achieve this processing strategy astronomers moved away from algebraic statistics towards algorithmic processes and swiftly changed the way that the field of astronomy did analysis. This change in process did not remove traditional statistics from astronomy, but it did create a whole new area of astronomy that is now part of core undergraduate astronomy courses. This change in the nature of analysis may well be the future of educational assessment, where vast quantities of data are produced every day. A future where algorithmic inference has a more substantial role in analysis.

The current study provides a glimpse into that possible future. The evidence produced here provides some evidence that algorithmic, rather than algebraic, models of assessment may be a viable alternative to current methods such as IRT and True Scores. It opens the possibility of utilizing large amounts of student temporal data to make inferences, rather than between-student data, which may alter the kinds of conclusions and categories that are developed. In some ways it shows an alternative starting place for the utilization of big data in education. There are many advocates for building models based on those currently common in education. For example developing standardized test technology further, building more complex IRT models, taking big data and reducing it to be interpreted in the ways that we are already familiar such as regression analysis. We are at a turning point in terms of methodology, we can build a new methodology on what made sense 100 years ago or can we start somewhere else. That is, we can use as a base samples and between-student differences or we can we start somewhere completely different, possibly where the individual over time plays a bigger role in our inferences.

Personalization

We can think of individualization as putting the most appropriate intervention in front of a particular student – and to some extent this is what it amounts to. In combination with big data it can take on a broader meaning: individualization is really an acknowledgement that the differences between people are far greater than we have been able to accommodate logistically, but now we may be able to.

In a world of experts and a strictly defined corpus of knowledge that can be held by a single person, it makes sense to standardize knowledge. It makes sense to ensure that everyone is held to that standard and that we measure people's ability to meet the standard. But individualization is suited to a different world, a world where the amount of knowledge, even with a well-defined field such as particle physics, is too expansive for a single person to retain. Instead of putting our energy into forcing everyone into the same standard of knowledge, we can accept that to a great degree people's knowledge will differ. It makes more sense, then, to spend energy trying to characterize that knowledge and its consequences, and training students to be able to do this for themselves.

To return to the snooker analogy, we have been training students to start the game, to *break* the triangle. Some students are very, very good at breaking and we can identify those students for the purpose of assessment. A broad interpretation of individualization in concert with big data is less concerned with this limited scope and more concerned with characterizing in detail student behavior. Feedback of this information to the student represents the possibility of education being less focused on the break and more focused on all the other plays during the game. Education might become about learning what the consequences of different actions might be two, three plays ahead rather than just performance on a single, well performed opening shot. For this possible future to occur we would require an ability to characterize how our understanding impacts our individual behavior and this hinges on being able to make sense of large data sets in real time and for this we likely require algorithmic processes such as the Inverse Bayes Filter.

References

Allais, M. (1953). The behaviour of the rational man faced with uncertainty. *Econometrica*, *21*(4), 503–46.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*(2), 167. doi:10.1207/s15327809jls0402_2

Ary, D., Jacobs, L., Sorensen, C., & Walker, D. (2013). *Introduction to research in education*. New York: Cengage Learning.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, *96*(1), 124–129. doi:10.1037/h0033475

Azevedo, R., Johnson, A., Chauncey, A., & Burkett, C. (2010). Self-regulated learning with MetaTutor: Advancing the science of learning with meta-cognitive tools. In M. S. Khine & I. M. Saleh (Eds.), *New Science of Learning* (pp. 225–247). Springer New York. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/978-1-4419-5716-0_11

Baddeley, M. C., Curtis, A., & Wood, R. (2004). An introduction to prior information derived from probabilistic judgements: elicitation of knowledge, cognitive bias and herding. *Geological Society, London, Special Publications*, *239*(1), 15–27. doi:10.1144/GSL.SP.2004.239.01.02

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. doi:10.1037/0033-295X.84.2.191

Bandura, A. (1993). Perceived Self-Efficacy in Cognitive Development and Functioning. *Educational Psychologist*, *28*(2), 117–148. doi:10.1207/s15326985ep2802_3

Benjamin, L. T. (1988). A history of teaching machines. *American Psychologist*, *43*(9), 703–712. doi:10.1037/0003-066X.43.9.703

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, *21*(1), 65–88. doi:10.1177/0146621697211006

Black, P., & Wiliam, D. (2006). *Inside the black box: Raising standards through classroom assessment*. Granada Learning.

Bloom, B. S. (1968). *Learning for mastery*. Regional Education Laboratory for the Carolinas and Virginia Durham, NC.

Borsboom, D. (2005). *Measuring the mind : conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 67–97). New York, NY: Springer New York. Retrieved from http://nrs.harvard.edu/urn-3:hul.ebookbatch.SPRGR_batch:9780387959214

Bouffard-Bouchard, T. (1990). Influence of self-efficacy on performance in a cognitive task. *The Journal of Social Psychology*, *130*(3), 353–363. doi:10.1080/00224545.1990.9924591

Bradbard, D. A., Parker, D. F., & Stone, G. L. (2004). An alternate multiple-choice scoring procedure in a macroeconomics course. *Decision Sciences Journal of Innovative Education*, *2*(1), 11–26. doi:10.1111/j.0011-7315.2004.00016.x

Bruning, R., Zygielbaum, A., & Grandgenett, N. (2001). Using online learning resources to promote deeper learning. *Teacher Education Faculty Proceedings & Presentations*. Retrieved from http://digitalcommons.unomaha.edu/tedfacproc/20

Cambridge Community Development Department. (2010). Demographics and Statistics FAQ: City of Cambridge, Massachusetts. Retrieved October 4, 2014, from http://www.cambridgema.gov/cdd/factsandmaps/demographicfaq.aspx

Choy, S. L., O'Leary, R., & Mengersen, K. (2009). Elicitation by design in ecology: Using expert opinion to inform priors for Bayesian statistical models. *Ecology*, *90*(1), 265–277.

Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement*, *16*(1), 13–37. doi:10.1177/001316445601600102

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278. doi:10.1007/BF01099821

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, *14*(1), 1–13.

Cronbach, L. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*(11), 671–684.

Cronbach, L., & Snow, R. (1981). *Aptitudes and instructional methods: A handbook for research on interactions*. New York: Irvington Publishers.

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, *18*(1), 87–123. doi:10.1111/j.2044-8317.1965.tb00695.x

Department of Education. (2010a). Individualized, Personalized, and Differentiated Instruction. Retrieved July 24, 2013, from https://www.ed.gov/technology/draft-netp-2010/individualized-personalized-differentiated-instruction

Department of Education. (2010b, September 2). Beyond the bubble tests: The next generation of assessments. Retrieved January 3, 2013, from https://www.ed.gov/news/speeches/beyond-bubble-tests-next-generation-assessments-secretary-arne-duncans-remarks-state-l

Des Jardins, S. L., & Toutkoushian, R. K. (2005). Are Students Really Rational? The Development of Rational Thought and its Application to Student Choice. In J. C. Smart (Ed.), *Higher Education: Handbook of Theory and Research* (pp. 191–240). Springer Netherlands. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/1-4020-3279-X_4

Desmarais, M. C., & Baker, R. S. J. d. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1-2), 9–38. doi:10.1007/s11257-011-9106-8

Douven, I. (1999). Inference to the best explanation made coherent. *Philosophy of Science*, *66*, S424–S435.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, *41*(10), 1040–1048. doi:10.1037/0003-066X.41.10.1040

Dweck, C. S. (2012). *Mindset*. London: Robinson.

Ebel, R. L. (1965). *Measuring educational achievement*. Englewood Cliffs, N.J.,: Prentice-Hall.

Echternacht, G. (1972). The use of confidence testing in objective tests. *Review of Educational Research*, *42*(2), 217–236.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, *75*(4), pp. 643–669.

Eser, Z., Holbrook, M. E., & Colbert, J. (2012). Confidence based marking: Implementation and feedback measures. *Journal of Higher Education Theory and Practice*, *12*(1), 27–38.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review*, *83*(1), 37–64. doi:10.1037/0033-295X.83.1.37

Falmagne, J.-C., Cosyn, E., Doignon, J.-P., & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In R. Missaoui & J. Schmidt (Eds.), *Formal Concept Analysis* (pp. 61–79). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/11671404_4

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307 – 314. doi:http://dx.doi.org/10.1016/j.tics.2004.05.002

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585. doi:10.1111/j.1467-9280.2009.02335.x

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education*, *2*(1), 79–96. doi:10.1207/s15324818ame0201_5

Friston, K. (2011). What Is optimal about motor control? *Neuron*, *72*(3), 488–498. doi:10.1016/j.neuron.2011.10.018

Gardner-Medwin, A. (2013, July 22). *Optimisation of certainty-based assessment scores*. Presented at the IUPS, Birmingham UK. Retrieved from http://www.ucl.ac.uk/~ucgbarg/tea/IUPS_2013a.pdf

Gardner-Medwin, A., & Gahan, M. (2003). Formative and summative confidence-based assessment. In *Proceedings of the 2008 International Computer Assisted Assessment (CAA) Conference* (pp. 147–155). London.

Gardner-Medwin, A. R. (1995). Confidence assessment in the teaching of basic science. *ALT-J*, *3*(1), 80–85. doi:10.1080/0968776950030113

Ghirardato, P. (2002). Revisiting Savage in a conditional world. *Economic Theory*, *20*(1), 83–92. doi:10.1007/s001990100188

Gopnik, A. (2008). Causal inference and counterfactual reasoning in scientists and children. *International Journal of Psychology*, *43*(3-4), 52–52.

Gopnik, A., & Glymour, C. (2002). Causal maps and Bayes nets: A cognitive and computational account of theory-formation. *The Cognitive Basis of Science*, 117–132.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 3–31.

Gopnik, A., & Tenenbaum, J. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science*, *10*(3), 281–287. doi:10.1111/j.1467-7687.2007.00584.x

Gordon, J. (2003). Assessing students' personal and professional development using portfolios and interviews. *Medical Education*, *37*(4), 335–340. doi:10.1046/j.1365-2923.2003.01475.x

Gottfries, N., & Hylton, K. (1987). Are M.I.T. students rational?: Report on a survey. *Journal of Economic Behavior & Organization*, *8*(1), 113 – 120. doi:http://dx.doi.org/10.1016/0167-2681(87)90024-2

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge University Press.

Griffiths, & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, *5*(11), 887–892. doi:10.1038/nrn1538

Grittner, F. M. (1975). Indivisualized instruction: An historical perspective. *The Modern Language Journal*, *59*(7), 323–333. doi:10.1111/j.1540-4781.1975.tb04708.x

Gupta, S. S. (1994). *Statistical decision theory and related Topics V*. New York, NY: Springer New York.

Harren, V. A. (1979). A model of career decision making for college students. *Journal of Vocational Behavior*, *14*(2), 119–133. doi:10.1016/0001-8791(79)90065-4

Jonassen, D. H., & Grabowski, B. L. (2012). *Handbook of individual differences learning and instruction*. Routledge.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. doi:10.1016/0010-0285(72)90016-3

Keller, F. S. (1974). Ten years of personalized instruction. *Teaching of Psychology*, *1*(1), 4–9.

Kelly, D., & Tangney, B. (2002). Incorporating Learning Characteristics into an Intelligent Tutor. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems* (pp. 729–738). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/3-540-47987-2_73

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review*, *19*(2), 195–229. doi:10.5465/AMR.1994.9410210745

Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology & Society*, *14*(4), 99–110.

Le Cam, L. (1955). An extension of Wald's theory of statistical decision functions. *The Annals of Mathematical Statistics*, *26*(1), pp. 69–81.

Leclercq, D. (1993). Validity, Reliability, and Acuity of Self-Assessment in Educational Testing. In P. D. A. Leclercq & P. J. E. Bruno (Eds.), *Item Banking: Interactive Testing and Self-Assessment* (pp. 114–131). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/978-3-642-58033-8_11

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*(2), 159–183. doi:10.1016/0030-5073(77)90001-0

Lindley, D. V. (2000). The Philosophy of Statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 293–337. doi:10.1111/1467-9884.00238

Lombardi, M. J., & Nicoletti, G. (2012). Bayesian prior elicitation in DSGE models: Macro- vs micropriors. *Journal of Economic Dynamics and Control*, *36*(2), 294–313. doi:10.1016/j.jedc.2011.09.010

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Pub. Co.

Mäkitalo-Siegl, K., & Fischer, F. (2011). Stretching the limits in help-seeking research: Theoretical, methodological, and technological advances. *Learning and Instruction*, *21*(2), 243–246. doi:10.1016/j.learninstruc.2010.07.002

Maragoudakis, M., Tselios, N. K., Fakotakis, N., & Avouris, N. M. (2002). Improving SMS usability using Bayesian networks. In I. P. Vlahavas & C. D. Spyropoulos (Eds.), *Methods and Applications of Artificial Intelligence* (pp. 179–190). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/3-540-46014-4_17

Marr, D., & Poggio, T. (1979). A Computational Theory of Human Stereo Vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *204*(1156), 301–328. doi:10.1098/rspb.1979.0029

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Matthews, R. A. J. (2001). Methods for assessing the credibility of clinical trial outcomes. *Drug Information Journal*, *35*(4), 1469–1478. doi:10.1177/009286150103500442

Mcquiggan, S. W., Mott, B. W., & Lester, J. C. (2008). Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction*, *18*(1), 81–123.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1037/0003-066X.50.9.741

Michael, J. J. (1968). The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement*, *5*(4), 307–314. doi:10.2307/1433782

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education*, *30*, 109–162.

Muraki, E. (1992). A generalized partial credit model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. doi:10.1177/014662169201600206

Neuringer, A. (1986). Can people behave "randomly?": The role of feedback. *Journal of Experimental Psychology: General*, *115*(1), 62–75. doi:10.1037/0096-3445.115.1.62

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Pajares, F., & David, M. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology*, *86*(2), 193–203. doi:10.1037/0022-0663.86.2.193

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis: A new alternative to knowledge tracing. In *The 14th International Conference on Artificial Intelligence in Education, 2009*. Brighton, UK.

Penfield, R. D., Myers, N. D., & Wolfe, E. W. (2008). Methods for assessing item, step, and threshold invariance in polytomous items following the partial credit model. *Educational and Psychological Measurement*, *68*(5), 717–733. doi:10.1177/0013164407312602

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302–321. doi:10.1016/j.cognition.2010.11.015

Pierce, C. S. (1877). Philosophical writings of Peirce, the fixation of belief. *Popular Science Monthly*.

Plano, R. J., & Toby, S. (2004). Testing, testing: Good teaching is difficult; so is meaningful testing. *Journal of Chemical Education*, *81*(2), 180. doi:10.1021/ed081p180

Popper, K. R. (1959). The propensity interpretation of probability. *The British Journal for the Philosophy of Science*, *10*(37), 25–42. doi:10.2307/685773

Ramsey, P. H., Ramsey, P. P., & Barnes, M. J. (1987). Effects of student confidence and item difficulty on test score gains due to answer changing. *Teaching of Psychology*, *14*(4), 206–210. doi:10.1207/s15328023top1404_3

Razzaq, L., & Heffernan, N. T. (2006). Scaffolding vs. hints in the Assistment system. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems* (pp. 635–644). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/11774303_63

Regian, J. W., Shute, V. J., & Shute, V. (2013). *Cognitive Approaches To Automated Instruction*. Routledge.

Rippey, R. (1968). Probabilistic Testing. *Journal of Educational Measurement*, *5*(3), 211–215. doi:10.2307/1433981

Sacks, J. (1963). Generalized Bayes solutions in estimation problems. *The Annals of Mathematical Statistics*, *34*(3), 751–768. doi:10.2307/2238460

Schlaifer, R., & Raiffa, H. (1961). *Applied statistical decision theory*. Boston, MA: Division of Research, Graduate School of Business Administration, Harvard University.

Seidensticker, R. B. (2006). *Future hype: the myths of technology change* (1st ed.). San Francisco, CA :Berkeley, CA: Berrett-Koehler Publishers ;Publishers Group West [distributor].

Shavelson, R. J., Roeser, R. W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., … Quihuis, G. (2002). Richard E. Snow's Remaking of the Concept of Aptitude and Multidimensional Test Validity: Introduction to the Special Issue, *8*, 77–99.

Shuford Jr, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika*, *31*(2), 125–145. doi:10.1007/BF02289503

Sia, D. T., Treagust, D. F., & Chandrasegaran, A. L. (2012). High school students' proficiency and confidence levels in displaying their understanding of basic electrolysis concepts. *International Journal of Science and Mathematics Education*, *10*(6), 1325–1345. doi:10.1007/s10763-012-9338-z

Siegler, R. S. (1996). *Emerging minds: the process of change in children's thinking.* New York: Oxford University Press.

Sinatra, G. M., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., … Corno, L. (2001). *Remaking the Concept of Aptitude: Extending the Legacy of Richard E. Snow.* Taylor & Francis.

Singpurwalla, N. D. (1992). A Bayesian perspective on Taguchi's approach to quality engineering and tolerance design. *IIE Transactions*, *24*(5), 18–31. doi:10.1080/07408179208964242

Skinner, N. (1983). Switching answers on multiple-choice questions: Shrewdness or shibboleth? *Teaching of Psychology*, *10*(4), 220–222. doi:10.1207/s15328023top1004_9

Snow, R. E. (1992). Aptitude Theory: Yesterday, Today, and Tomorrow. *Educational Psychologist*, *27*(1), 5–32. doi:10.1207/s15326985ep2701_3

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology*, *76*(3), 347–376. doi:10.1037/0022-0663.76.3.347

Speece, D. L. (1990). Aptitude-Treatment Interactions: Bad rap or bad idea? *The Journal of Special Education*, *24*(2), 139–149. doi:10.1177/002246699002400203

Stein, C. (1955). A necessary and sufficient condition for admissibility. *The Annals of Mathematical Statistics*, *26*(3), 518–522.

Sullivan, A. (2006). Students as rational decision-makers: the question of beliefs and attitudes. *London Review of Education*, *4*(3), 271–290. doi:10.1080/14748460601043965

Swinton, S. S. (1987). The predictive validity of the restructured GRE with particular attention to older students. *ETS Research Report Series*, *1987*(1), i–18. doi:10.1002/j.2330-8516.1987.tb00226.x

Tan, H., Wang, J., Shen, C., & Xu, Y. (2005). Trajectory programming algorithm for cruise missile guidance. *Journal of Nanjing University of Aeronautics & Astronautics*, *1*.

Taylor, C., & Gardner, P. L. (1999). An alternative method of answering and scoring multiple choice tests. *Research in Science Education*, *29*(3), 353–363. doi:10.1007/BF02461598

Tenenbaum. (2000). Rules and Similarity in Concept Learning. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems* (p. 1098). MIT Press.

Tenenbaum, J. B. (1996). Learning the structure of similarity. *Advances in Neural Information Processing Systems*, 3–9.

Tian, G.-L., Ng, K. W., & Tan, M. T. (2010). *Bayesian missing data problems : EM, data augmentation and noniterative computation*. Boca Raton: Chapman & Hall/CRC.

Tian, G.-L., & Tan, M. (2003). Exact statistical solutions using the Inverse Bayes Formulae. *Statistics & Probability Letters*, *62*(3), 305–315. doi:10.1016/S0167-7152(03)00044-0

Tian, G.-L., Tan, M., & Ng, K. W. (2007). An exact non-iterative sampling procedure for discrete missing data problems. *Statistica Neerlandica*, *61*(2), 232–242. doi:10.1111/j.1467-9574.2007.00345.x

Tibbetts, S. G. (1997). Gender differences in students' rational decisions to cheat. *Deviant Behavior*, *18*(4), 393–414. doi:10.1080/01639625.1997.9968068

Vonderwell, S. K., & Boboc, M. (2013). Promoting formative assessment in online teaching and learning. *TechTrends*, *57*(4), 22–27. doi:10.1007/s11528-013-0673-x

Wald, A. (1950). *Statistical decision functions*. Oxford, England: Wiley.

Wang, Y., & Heffernan, N. (2010). Representing student performance with partial credit. In *Educational Data Mining 2010, The 3rd International Conference on Educational Data Mining*. Pittsburgh.

Wang, Y., & Heffernan, N. (2011). *Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes*. Retrieved from http://web.cs.wpi.edu/~nth/pubs_and_grants/papers/2013/AIED2013/YuTaoContinousNodeSub.pdf

Wang, Y., & Heffernan, N. (2013). Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial Intelligence in Education* (pp. 181–188). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/chapter/10.1007/978-3-642-39112-5_19

Winkler, R. L. (1967). The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, *62*(319), 776–800. doi:10.2307/2283671

Woodworth, R. S. (Ed.). (1915). *Archives of Psychology* (Vol. 4). New York: The Science Press.

Yeh, Y. (2012). Aptitude-Treatment Interaction. In P. D. N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 295–298). Springer US. Retrieved from http://link.springer.com.ezp-prod1.hul.harvard.edu/referenceworkentry/10.1007/978-1-4419-1428-6_582

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian Knowledge Tracing models. In *Artificial Intelligence in Education* (pp. 171–180). Springer.

VITA

Charles William McLeod Lang

2001 – 2007	The University of Melbourne Victoria, Australia	B.A. March 2007
		B.Sc. March 2007
2007 – 2008	Harvard Graduate School of Education Cambridge, MA	Ed.M. May 2008