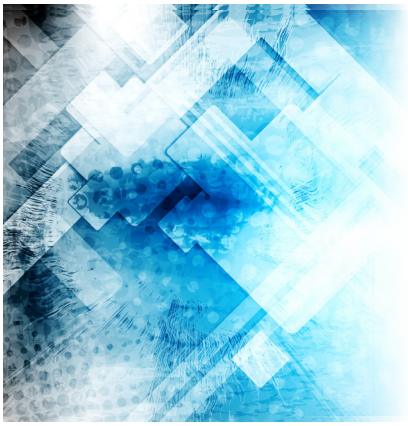
RESEARCH & PRACTICE IN ASSESSMENT



AUTHORS
Ryan S. Baker, Ph.D.
Teachers College, Columbia University

Albert T. Corbett, Ph.D.
Carnegie Mellon University

Abstract

Many university leaders and faculty have the goal of promoting learning that connects across domains and prepares students with skills for their whole lives. However, as assessment emerges in higher education, many assessments focus on knowledge and skills that are specific to a single domain. Reworking assessment in higher education to focus on more robust learning is an important step towards making assessment match the goals of the context where it is being applied. In particular, assessment should focus on whether learning is robust (Koedinger, Corbett, & Perfetti, 2012), whether learning occurs in a way that transfers, prepares students for future learning, and is retained over time; and also on skills and meta–competencies that generalize across domains. By doing so, we can measure the outcomes that we as educators want to create, and increase the chance that our assessments help us to improve the outcomes we wish to create. In this article, we discuss and compare both traditional test–based methods for assessing robust learning, and new ways of inferring robustness of learning while the learning itself is occurring, comparing the methods within the domain of college genetics.

Assessment of Robust Learning with Educational Data Mining

In recent years, the historical monopoly of universities in higher education has been challenged by new entrants, including for–profit universities and massive online open courses (Hanna, 1998; Vardi, 2012). This change has brought to the forefront questions about what the core goals of higher education are: Is it to train a workforce in specific employable skills (Sperling & Tucker, 1997)? Or is it to promote learning that connects across domains and prepares students to learn the new skills and disciplines that emerge during their years in the workforce (Knapper & Croppley, 2000)? To put it another way, is the goal of higher education to learn competencies, or to learn meta–competencies which cut across domains (e.g., Buckingham Shum & Deakin Crick, 2012)?

While much of the learning that goes on in higher education pertains primarily to the content area of the class being taken, students can learn in a specific fashion or in a more general fashion. Increasingly, researchers in the learning sciences have presented evidence that it is possible to measure whether learning is *robust* – defined in Koedinger et al. (2012) as learning that can transfer to related situations (Fong & Nisbett, 1991; Singley & Anderson, 1989), prepares students for future learning (Bransford & Schwartz, 1999; Schwartz & Martin, 2004), and is retained over the long–term (Bahrick, Bahrick, Bahrick & Bahrick, 1993; Schmidt & Bjork, 1992).

CORRESPONDENCE

Email
ryanshaunbaker@gmail.com

To the extent that creating more robust learning is the primary goal of higher education, the way assessment is used may need to change. While some argue for a switch to self–assessment (e.g., Boud & Falchikov, 2006), we still see a need for instructor and curriculum–led assessment. But there is a challenge for those developing assessments for higher education; it is much easier to measure didactic knowledge or concrete skill than to measure the type of learning that has been argued for.

Nonetheless, whether learning is robust can be measured. Paper tests measuring retention and transfer have been in use for quite some time (cf. Gick & Holyoak, 1983; Surber & Anderson, 1975), with paper tests measuring a student's preparation for future learning (PFL) emerging about a decade ago (Bransford & Schwartz, 1999; Schwartz &

Martin, 2004). In this article, we discuss examples of this work within the domain of college genetics. Increasingly, it is also a goal of assessment in higher education to measure skills that cut across domains, such as science inquiry and help seeking (cf. Puncochar & Klett, 2013), and to measure robust learning of these skills while learning is ongoing (cf. Linn & Chiu, 2011). To this end, we will also discuss measures of robust learning that can measure robust learning of domain content, but also domain–general skills, in a fashion that is integrated into instruction. We discuss these new forms of assessment in terms of the same domain of college genetics for understandability; but, as we will discuss, many of the new forms of assessment are potentially meaningful domain–general.

These new forms of assessment are based on the emerging methods of educational data mining (EDM; Baker & Siemens, in press; Baker & Yacef, 2009; Romero & Ventura, 2007). Within educational data mining, the voluminous data increasingly becoming available to learners, particularly from online learning environments, becomes a source of information that can be used to identify complex learning behaviors and ill–defined or complex skill (cf. Kinnebrew & Biswas, 2012; Sao Pedro, Baker, & Gobert, 2012). These data are sometimes analyzed by use of knowledge engineering methods, where research analysts identify meaningful patterns in data by hand (e.g., Aleven, McLaren, Roll, & Koedinger, 2006), and is sometimes analyzed using automated methods such as sequence mining (Kinnebrew & Biswas, 2012) or classification (Sao Pedro et al., 2012). While knowledge engineering can be similar to traditional psychometric approaches for assessment development such as evidence–centered design (Mislevy, Almond, & Lukas, 2004), and advanced ECD–based models of complex student skill can resemble EDM models developed using automated discovery (see Shute & Ventura, 2013 for examples), the development methods of EDM and ECD differ, as do their validation. Educational data mining methods are often validated by developing the models on one set of students and testing them on another; some EDM methods are also validated on data from new domains or contexts (Sao Pedro, Gobert, & Baker, 2014) or data from new learner populations (Ocumpaugh, Baker, Kamarainen, & Metcalf, 2014). In addition, EDM–based assessments are typically validated for agreement with human judgments about a construct's presence which themselves are known to be reliable (Ocumpaugh et al., 2014; Sao Pedro et al., 2014), and are based on data features thought by domain experts to be plausibly related to the construct of interest (Sao Pedro et al., 2012). In some cases, their internal structure is not considered in detail, being too complex for a human analyst to understand without hours of study, but that is not true of all EDM–developed models; the models resulting from the EDM process are particularly simple for the cases presented in this paper. A full discussion of educational data mining methods is outside the scope of this paper, but richer summaries are provided in the papers (Baker & Siemens, in press; Baker & Yacef, 2009; O'Reilly & Veeramachaneni, 2014; Romero & Ventura, 2007) and the textbook (Baker, 2013).

To put it another way, is the goal of higher education to learn competencies, or to learn meta–competencies which cut across domains?

EDM–based assessment has multiple benefits compared to traditional methods of assessment: If the models are designed appropriately, they can be used in real time to make assessment during learning and support real time intervention. In addition, since the models typically make inferences based on ongoing interaction between a student and online system, they can replicate the assessments made by more traditional instruments without needing to take the student's time up with a paper test. See, for instance, Feng, Heffernan, and Koedinger (2009), who show that EDM models based on student interaction can accurately predict standardized exam scores.

Case Study in College Genetics Tutor

In this article, we discuss the potential for assessment of robust learning in higher education, both with traditional methods and educational data mining methods, using examples drawn from the domain of genetics. Genetics is an important topic because it is a central, unifying theme of modern biology and because it provides the foundation for many advances in 21st century technology. It is a challenging topic for students, because it depends heavily on problem solving (Smith, 1988). Finally, it is a relevant topic because it affords an interesting form of superficial learning: Students can develop successful problem solving algorithms that are not well grounded in the underlying biology.

We discuss this specifically within the context of work to develop and utilize an e–learning system for college genetics, the Genetics Cognitive Tutor (GCT; Corbett, Kauffman,

MacLaren, Wagner, & Jones, 2010). GCT is focused on helping students learn not only genetics domain materials, but also the complex abductive reasoning skills needed to make inferences within this domain. Abductive reasoning skills involve reasoning "backward" from empirical observations (e.g., a daughter of unaffected parents is affected by a genetic trait) to an explanation for the observations (each parent must carry a recessive allele for the trait). Abductive reasoning skills are an important part of the undergraduate learning experience, not just in genetics, but across domains, because they are essential skills in formulating scientific knowledge, and in applying such knowledge to diagnostic tasks.

Cognitive Tutors are a type of online learning system where students complete problems (in genetics or other domains) within the context of activities designed to scaffold problem solving skill (Koedinger & Corbett, 2006). The student completes problems within an interface that makes visible cognitive steps of the problem solving process visible, and receives instant feedback on their performance. Student performance is analyzed in real time according to a cognitive model of the domain. If a student's answer indicates a known misconception, the student receives instant feedback on why their answer was incorrect. At any time, the student can request help that is sensitive to their current learning context.

To the extent that creating more robust learning is the primary goal of higher education, the way assessment is used may need to change.

GCT has more than 175 genetics problems, divided into 19 modules, which address topics in Mendelian inheritance, pedigree analysis, genetic mapping, gene regulation, and population genetics. An average of about 25 steps is needed for each of the 175 problems in GCT. It has served as supplementary instruction in a variety of undergraduate biology classes in a wide range of public and private universities in the United States and Canada (Corbett et al., 2010). It has also been used by students enrolled in high school biology classes (e.g., Corbett et al., 2013a, 2013b; Baker, Corbett, & Gowda, in press).

The goal of GCT is not just to promote immediate learning of the exact content studied within the system, but to promote robust learning as defined above. As such, research during the development of GCT focused on assessing robust learning, both after use of the system and during use of the system.

Assessing Robust Learning in College Genetics with Tests

Tests historically have been one of the most common methods for assessing robust learning. They are clearly the most straightforward way of doing so; for instance, a test can be administered immediately at the end of an activity or multiple times during the semester.

Within educational data mining, the voluminous data increasingly becoming available to learners, particularly from online learning environments, becomes a source of information that can be used to identify complex learning behaviors and ill–defined or complex skill.

The history of research on retention of material, both in research settings and classroom settings, has depended heavily on retesting the same material or same skill. This has been conducted through classical paper tests (Surber & Anderson, 1975), and in online systems such as the Automatic Reassessment and Relearning System, which retests a student on material they have learned at increasing time intervals (Wang & Heffernan, 2011).

So too, a great deal of the research on whether knowledge is transferrable has depended on paper tests, although performance–based measures have also been used in some cases (e.g., Singley & Anderson, 1989). And again, while much of the research on preparation for future learning has utilized complex learning activities and resources, the assessments have often involved paper post–tests, albeit post–tests with learning resources embedded (e.g., Bransford & Schwartz, 1999; Chin et al., 2010, Schwartz & Martin, 2004).

In several GCT studies, paper assessments of retention, transfer, and PFL were administered to study the robustness of student learning. For a selected set of lessons, transfer tests and PFL tests were administered to students immediately after they completed use of the system. For example, after students completed a lesson on 3–factor cross reasoning, they were assigned "gap filling transfer tests" (VanLehn, Jones, & Chi, 1992) where they had to complete problems for which a core case in the original formulas they learned did not apply. The problem is solvable and most of the students' problem solving knowledge directly applies; however, the student can only complete the task if they can draw on their conceptual understanding of that problem solving knowledge to fill in the gap that results from the missing group.

In the preparation for future learning tests, material beyond the current lesson was involved. For example, for the PFL test for a lesson on 3–factor cross, students were asked to solve parts of a more complex 4–factor cross problem. The reasoning is related to solving a 3–

factor cross problem, but substantially more complicated, making it unlikely that the student could discover an effective solution method during the test. Instead, the test gave the student a textual description of the solution method, and then asked them to solve the problem. For retention, the same types of problems as seen in GCT were given to students in a paper form, but one week later.

Students were generally successful on each of these tests. Student performance on the test of retention was high (M = 0.78, SD = 0.21), comparable to the immediate post–test that covered the same skills as the lesson (M = 0.81, SD = 0.18), and substantially higher than the pre–test (M = 0.31, SD = 0.18). Student performance on the PFL test (M = 0.89, SD = 0.15) and transfer test (M = 0.85, SD = 0.18) was also high, approximately equal to the immediate basic problem–solving post–test (Baker, Gowda, & Corbett, 2011a, 2011b). These results indicated that the GCT was generally successful at promoting robust learning.

It would be possible to stop at this point, and simply offer that conclusion; however, it would be useful to be able to infer the robustness of student learning earlier than after the learning episode. Beyond that, it is desirable to be able to infer the robustness of learning during the learning episode, when it is easier to intervene. In addition, tests are time consuming to administer. As such, the following sections describe our work to infer robust learning in real time, and thus these tests were used as the basis for further research.

Inferring Robust Learning in College Genetics with Learning Models

A second way to infer robust learning is through the use of automated models that infer student skill learning. This method is not specifically tailored to robust learning – it is tailored to the learning that occurs in the lesson being studied – but may be successful at predicting robust learning as well. There are examples of this type of research going back several years. For example, Jastrzembski, Gluck, and Gunzelmann (2006) have used this type of modeling to predict student retention of knowledge, within an online learning system teaching flight skills.

In this article, we discuss the potential for assessment of robust learning in higher education, both with traditional methods and educational data mining methods, using examples drawn from the domain of genetics.

Within GCT, knowledge is modeled in real time using an algorithm named Bayesian Knowledge Tracing (Corbett & Anderson, 1995). Bayesian Knowledge Tracing (BKT) is the classic algorithm for modeling student knowledge within online problem solving; it has been used in many systems and analyses, cited thousands of times, and performs comparably to or better than other algorithms for cases where its assumptions apply (see results and review in Pardos, Baker, Gowda, & Heffernan, 2011).

Bayesian Knowledge Tracing can be seen as either a simple Bayes Net or a simple Hidden Markov Model (Reye, 2004). Within BKT, a probability is continually estimated for the probability that the student knows each skill in the lesson or system. These probabilities are updated each time a student attempts a new problem solving step, with correct actions treated as evidence the student knows the skill, and incorrect actions and help requests treated as evidence that the student does not know the skill. As with psychometric models such as DINA (deterministic inputs, noisy and gate; Junker & Sijtsma, 2001), (Junker & Sijtsma, 2001), BKT takes into account the possibility that a student may have gotten a correct answer by guessing, or may have slipped and obtained an incorrect answer despite knowing the relevant skill. However, BKT does not typically account for the possibility that a student may forget what they have learned (but see an example where it is extended to do so in Qiu, Qi, Lu, Pardos, & Heffernan, 2011), or that a student may have developed shallow knowledge that will not transfer between contexts.

Bayesian Knowledge Tracing and its properties are discussed in detail in dozens of papers, with the first being Corbett and Anderson (1995). For reasons of space, only a brief description will be given here. Bayesian Knowledge Tracing calculates the probability that a student knows a specific skill at a specific time, applying four parameters within a set of equations, and repeatedly updating probability estimates based on the student's performance. This process is carried out separately for each of the cognitive skills in the domain – there are eight such skills in the case of the GCT lesson on 3–factor cross. The model makes the assumption that at each problem step, a student either knows the skill or does not know the skill. It was originally thought that the model also made the assumption that each student response will either be correct or incorrect (help requests are treated as incorrect by the model), but it has been shown more recently that extending BKT to handle probabilistic input

is very easy (e.g., Sao Pedro et al., 2014). If the student does not know a specific skill, there is nonetheless a probability **G** (for "Guess") that the student answer correctly. Correspondingly, if the student does not know the skill, there is a probability **S** (for "Slip") that the student will answer incorrectly. When the student starts the lesson, each student has an initial prior probability $\mathbf{L}_0$ of knowing each skill, and each time the student encounters the skill, there is a probability **T** (for "Transition") that the student will learn the skill, whether or not they answer correctly. Each of the four parameters within Bayesian Knowledge Tracing are fit for each skill, using data on student performance; there is current debate on which method is best for fitting parameters, but several approaches seem reasonable and comparably good (see discussion in Pardos et al., 2011).

Every time the student attempts a problem step for the first time, BKT updates its estimate that the student knows the relevant skill. The procedure is as follows (the relevant equations are given in Figure 1):

1.) Take the probability that the student knew the skill before the current problem step $\mathbf{L}_{n-1}$ and the correctness of the student response, and re–estimate the probability that the student knew the skill before the current problem step.

2.) Estimate the probability that the student knows the skill after the current problem step, using the adjusted probability that the student knew the skill before the current problem step, and the probability **T** that the student learned the skill on the step.

$$P(L_{n-1}|Correct_n) = \frac{P(L_{n-1}) * (1 - P(S))}{P(L_{n-1}) * (1 - P(S)) + (1 - P(L_{n-1})) * (P(G))}$$

$$P(L_{n-1}|Incorrect_n) = \frac{P(L_{n-1}) * P(S)}{P(L_{n-1}) * P(S) + (1 - P(L_{n-1})) * (1 - P(G))}$$

$$P(L_n|Action_n) = P(L_{n-1}|Action_n) + ((1 - P(L_{n-1}|Action_n)) * P(T))$$

Figure 1. The equations used to infer student latent knowledge from performance in Bayesian Knowledge Tracing.

Abductive reasoning skills are an important part of the undergraduate learning experience, not just in genetics, but across domains, because they are essential skills in formulating scientific knowledge, and in applying such knowledge to diagnostic tasks.

BKT, when applied to data from the GCT, was moderately successful at predicting transfer, PFL, and retention test performance (Baker et al., 2011a, 2011b; Baker et al., in press). By the end of the student's use of the tutor, BKT could achieve a correlation of 0.353 to transfer for new students, a correlation of 0.285 to PFL for new students, and a correlation of 0.305 to retention for new students. These levels of agreement were clearly better than no agreement, but still far from perfect. However, one positive for this method is that BKT–based predictions of robust learning were able to achieve close to this level of performance with only a subset of the data (the first 30% in the case of transfer). The performance of the BKT model at predicting transfer, as the student completes increasing amounts of the activity, is shown in Figure 2. In other words, the full degree of predictive power available from this method becomes available when the student has 70% more of the activity to complete. Even when prediction is imperfect, it can still be useful for intervention and automated adaptation if it is available early in the learning process.

Inferring Robust Learning in College Genetics with Meta–cognitive Behaviors

In order to improve upon these models, we next distilled features of the students' interaction with GCT that indicated student behaviors relevant to their meta–cognition. As robust learning involves more complex reasoning about material and conceptual understanding than simply whether the student can obtain the correct answer or not, we analyzed some of

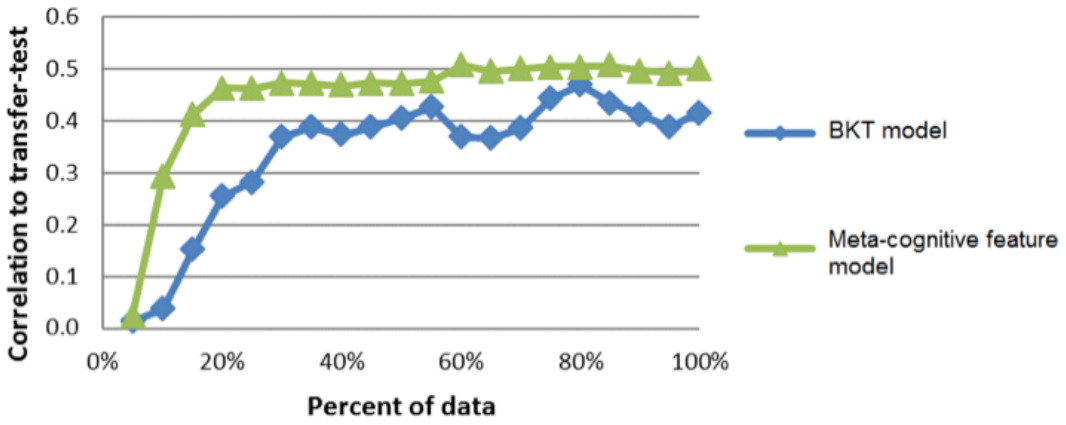


Figure 2. Predicting transfer with first N percent of the data. Graph reproduced with minor modifications from Baker et al. (2011a).

the more complex aspects of student behavior during learning. In doing so, we focused on behaviors that were informative about whether the student was demonstrating meta–cognition, and their engagement with the material. An example of such behavior might be when the software indicates to the student that their response involves a known misconception, and explains why the student's answer was wrong. Does the student pause to think through this explanation, or do they hurry forward without thinking carefully?

The history of research on retention of material, both in research settings and classroom settings, has depended heavily on retesting the same material (or same skill).

A set of 18 features reflective of student thinking were distilled from the students' interactions with the learning system, as shown in Table 1. As also shown in the table, several of these features were found to be individually predictive of PFL and transfer among college students (Baker et al., in press), but only one feature was predictive of retention. When combined into an integrated model (which used some but not all of these features, as some did not provide additional predictive power once other features were incorporated), all three models relied on whether the student sought help when they were struggling, or avoided help. The PFL model also relied upon whether the student paused to self–explain the hints they received. In addition to help seeking, the transfer model relied on whether students made fast actions that did not involve gaming the system (trying to get through the material without learning, for example by systematically guessing; cf. Baker, Corbett, Koedinger, & Wagner, 2004).

This produced the following models of transfer, PFL, and retention:

Transfer = – 1.5613 * HelpAvoidance(1) + 0.2968 * FastNotGaming(7') + 0.8272

PFL = 0.0127 * Spikiness(9) – 0.5499 * HelpAvoidance(1) – 5.3898 * LongPauseAfterHint(4) + 0.8773

Retention = – 2.398 * HelpAvoidance (1) + 0.852

When applied to new students, the transfer model achieved a correlation of 0.396 (Baker et al., in press), the PFL model achieved a correlation of 0.454 (Baker et al., in press), and the retention model achieved a correlation of 0.410. As such, model performance was better than using BKT estimates of student knowledge alone, although only moderately so. By contrast, the models of retention based on these features did not improve on the knowledge–based models.

In addition, these predictions of robust learning were able to achieve nearly this level of performance with only a subset of the data (the first 20% in the case of transfer), moderately faster than the knowledge–based models. In other words, the full degree of predictive power available from this method becomes available when the student has 80% more of the activity to complete, giving plenty of time for interventions designed to improve the robustness of learning. The performance of the meta–cognitive model at predicting transfer, as the student completes increasing amounts of the activity, is shown in Figure 2.

…we focused on behaviors that were informative about whether the student was demonstrating meta–cognition, and their engagement with the material. An example of such behavior might be when the software indicates to the student that their response involves a known misconception, and explains why the student's answer was wrong. Does the student pause to think through this explanation, or do they hurry forward without thinking carefully?

Table 1	
Features Distilled for Development of Automated Detectors of Robust Learning	
1	Help avoidance: the failure to request help on a skill the student does not know, on their first attempt (Aleven et al., 2006) *%&
1'	Not requesting help on skills the student already knows
2	An extended pause after receiving feedback for a known misconception *
2'	A short pause after receiving feedback for a known misconception *
3	An extended pause after receiving on-demand help messages *%
3'	A short pause after receiving on-demand help messages %
4	An extended pause after receiving on-demand help messages and getting the answer right (Shih, Koedinger, & Scheines, 2008) *%
4'	A short pause after receiving on-demand help messages and getting the answer right %
5	Long pauses on skills that the student probably knows *%
5'	Short pauses on skills that the student probably knows %
6	Off-task behavior, where the student is not working with the system or learning domain for an extended period of time (assessed using automated detector from Baker, 2007)
6'	Long pauses that are not assessed by the detector as off-task
7	Gaming the system, attempting to complete problems without learning the material, for example by systematically guessing or clicking rapidly through hints to get the answer (assessed using automated detector from Baker, Corbett, Roll, & Koedinger, 2008) *%
7'	Fast actions that do not involve gaming the system *%
8	The student's average probability of careless errors, making an error when the student is thought to have obtained the relevant skill (assessed using automated detector from Baker et al., 2010)
8'	The model's average certainty that a careless error is careless, e.g. the average of this construct when the probability is over 0.5
9	The student's average learning per problem step, according to the moment-by-moment learning model (Baker, Goldstein, & Heffernan, 2011) %
9'	The spikiness of the moment-by-moment learning model, e.g. the ratio between the maximum moment-by-moment learning and the average moment-by-moment learning (Baker et al., 2011) %*

Note. Greater operational detail on features is given in (Baker et al., in press). Features predictive of PFL are marked with a *. Features predictive of transfer are marked with a %. Features predictive of retention are marked with a &.

It is useful to know that these measures of meta–cognitive skill are predictive of robust learning in the domain of genetics. However, these measures are potentially applicable at greater scale than simply a single domain. For instance, the help seeking, help avoidance, and self–explanation models used in this analysis were originally developed in the context of mathematics (e.g., Aleven et al., 2006; Shih et al., 2008). In these previous papers, these same three models were shown to correlate to student learning outcomes. As the exact same models can predict learning outcomes both in high school mathematics and in college genetics, our current results – in combination with the previous results published by other authors – suggest that these models may capture aspects of learning skill that are domain–general. An important next step would be to see if these models' predictions are accurate, for the same student, in new domains. Showing that a model predicts learning outcomes in two domains is different than showing that a student's skill is domain general. In one example of this type of research, Sao Pedro and colleagues (2014) found that students who demonstrate scientific inquiry skill in one science domain are likely to be able to demonstrate the same skill in another domain.

Inferring Robust Learning in College Genetics with Moment–by–Moment Learning Models

A third method for inferring robust learning in college genetics that was tried is moment–by–moment learning models. The moment–by–moment learning model (Baker et al., 2011) is a distillate of Bayesian Knowledge Tracing that tries to infer not just the probability that a student has learned a skill by a certain point in a learning activity, but how much they learned at that stage of the activity. This inference is made using a combination of their current estimated knowledge, their behavior during the current learning opportunity, and their performance in the learning opportunities immediate afterwards.

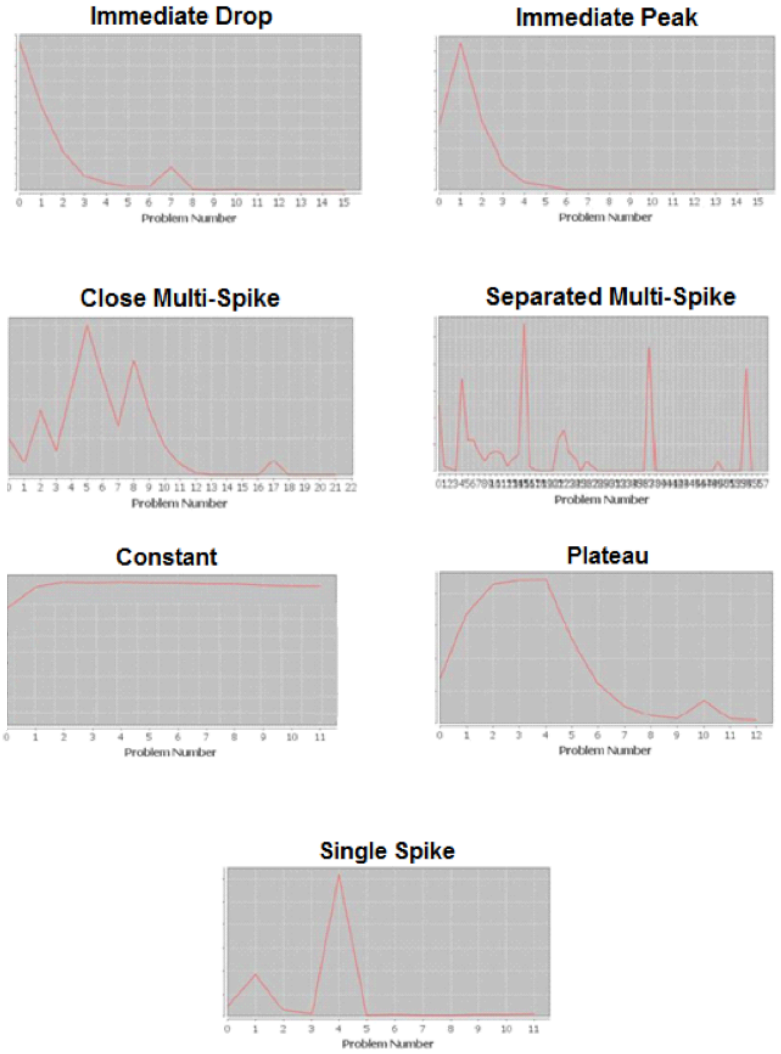


Figure 3. Examples of the visual features of moment–by–moment learning graphs studied by data coders. The x–axis on these graphs represents the number of problems or problem steps where the student has encoundered a specific skill; the y–axis represents the amount of learning inferred to have occurred during the problem step, relative to other problem steps. Note that these graphs show relative differences in learning rather than absolute amounts of learning, in order to facilitate visual interpretation by coders. Graphs reproduced from Baker et al. (2013).

The full mathematical details of this model are outside the scope of this paper and take up multiple pages, but are given in full in Baker et al.'s (2011) work. In brief, a combination of the probability of knowledge at the current time (according to BKT) is combined with data on the next two actions, in order to assess the probability of three cases at each time point: The student already knew the skill, the student did not know it but learned it at that time, and the student did not know the skill and did not learn it. Then, machine learning is used to smooth the inferences with additional data on student behavior, including help seeking and pauses. The details of the exact model used to do this smoothing in the case of genetics are given in Baker, Hershkovitz, Rossi, Goldstein, and Gowda's (2013) work.

Visual analysis of moment–by–moment learning over time indicated that there can be very different patterns in different students' learning, or in the learning of the same student for different skills (Baker et al., 2013). Examples are shown in Figure 3. One intuition was that certain patterns during the process of learning may indicate more or less robust learning. This intuition was supported by analyses where human coders labeled graphs by hand in terms of specific patterns, such as plateaus, hillsides, or single–spike graphs, and then these patterns were correlated to robust learning outcomes in GCT (Baker et al., 2013). Examples of these graphs are shown in Figure 3. Some patterns such as plateaus appeared to be correlated to less

In other words, the full degree of predictive power available from this method becomes available when the student has 80% more of the activity to complete, giving plenty of time for interventions designed to improve the robustness of learning.

robust learning, whereas other patterns such as hillsides, where the student learns the skill quickly upon beginning to use the system, appeared to be correlated to more robust learning. These patterns generally held across all three forms of robust learning.

Next, attempts were made to automate this process, distilling mathematical features of the graphs of learning over time, and building these into models to predict robust learning automatically within GCT (Hershkovitz, Baker, Gowda, & Corbett, 2013). The best model of PFL involved the area under the graph (an indicator of total learning), the height of the third–largest peak (the problem step where the third–most learning occurred), and the relative differences both in magnitude and time between the largest peak and the third–largest peak. This model achieved a correlation to PFL of 0.532 for new students, a better performance than the models based on meta–cognitive behaviors or knowledge. This work has not yet been replicated for transfer or retention. However, this model has one disadvantage compared to those models. Although it does not require the application of time consuming post–tests, it cannot infer the robustness of student learning until the student has completed the learning activity, making it less useful for immediate intervention during learning.

Conclusion

In this article, we have discussed multiple ways that robust learning can be inferred within higher education. One popular option is post–tests, whether administered online or on paper. For summative purposes, tests are likely to remain the gold standard option for some time. However, the data from online learning, in combination with educational data mining, provides an alternative with some benefits. Post–tests are time consuming to administer, and cannot be given in real time (particularly for retention tests, which by definition must be administered at a considerable delay). Models that can infer and predict robust learning from learning process data can make predictions which correlate to student robust learning outcomes, predictions which are available to instructors and for personalization within online learning systems much more quickly than paper tests can be available. At some cost to predictive power, predictions can be available as early as when the student has completed only 20% of the learning task. They can also help us to better understand the processes which lead to robust learning.

In our work with the Genetics Cognitive Tutor, we have developed three approaches to inferring robust learning: knowledge–based modeling, metacognitive–behavior–based modeling, and moment–by–moment–learning–based modeling.

In our work with the Genetics Cognitive Tutor, we have developed three approaches to inferring robust learning: knowledge–based modeling, metacognitive–behavior–based modeling, and moment–by–moment–learning–based modeling. The knowledge–based modeling approach was simplest to create as it depended solely on a standard model for measuring learning in online problem–solving; its performance was, however, the weakest. The approach based on modeling metacognitive behaviors required more effort to create; it reached asymptotic performance at inferring transfer and PFL after the student had completed 20% of the learning activity. Finally, the approach based on the moment–by–moment–learning–model was best at inferring PFL, but is not applicable until the student has completed the learning activity.

As such, models like the meta–cognitive behavior model are probably most relevant for use in automated interventions that attempt to infer which students are at risk of developing shallow learning and intervene in real time to enhance their learning. By contrast, models like the moment–by–moment–learning model are probably most relevant for informing instructors after an activity in which students have not developed robust learning, or for recommending additional alternate activities after a student completes an activity without achieving robust learning. Either approach is more work during development than simply creating a test; but these approaches have the potential to speed up assessment and facilitate giving students more rapid learning support.

Beyond their ability to predict tests of robust learning in a specific domain, these types of new measures may point the way to new domain–general assessment of student skills. In particular, the types of help seeking skills used in the meta–cognitive model have the potential to be domain–general, as science inquiry skills have been shown to be (e.g., Sao Pedro et al., 2014). It is not yet clear whether the moment–by–moment learning model indicators of robust learning will also prove general, but this is a valuable potential area for future work.

The importance of robust learning for higher education is clear. The goal of an undergraduate education is not simply to produce mastery of a known set of skills, or awareness of a known set of knowledge, but to prepare students for their future careers, where they will

have to be able to transfer their knowledge to new situations and contexts, and where they will need to be prepared for future learning, both in the domains they have studied and in the new areas that will emerge after they complete their studies.

As such, it is important to assess robust learning in higher education, and to support students in developing it. The approaches presented here represent a variety of ways that may make assessment of robust learning more feasible in the higher education context.

AUTHOR'S NOTE

We would like to thank Jaclyn Ocumpaugh for assistance in the early stages of conceptualizing and preparing this article, Annie Levy for assistance in the technical aspects of paper preparation, and the anonymous reviewers for helpful comments and suggestions. We would like to acknowledge Award # DRL–091018 from the National Science Foundation.

References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). *Toward meta–cognitive tutoring: A model of help seeking with a Cognitive Tutor. International Journal of Artificial Intelligence and Education*, *16*, 101–128.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4*, 316–321.

Baker, R. S. J. d. (2007). Modeling and understanding students' off–task behavior in intelligent tutoring systems. *Proceedings of ACM CHI 2007: Computer–Human Interaction*, 1059–1068.

Baker, R. S. J. d. (2013). Learning, schooling, and data Analytics. In M. Murphy, S. Redding, & J. Twyman (Eds.), *Handbook on innovations in learning for states, districts, and schools* (pp.179–190). Philadelphia, PA: Center on Innovations in Learning.

Baker, R. S. J. d., Corbett, A. T., & Gowda, S. M. (in press). Generalizing automated detection of the robustness of student learning in an intelligent tutor for genetics. *Journal of Educational Psychology*.

Baker, R. S. J. d., Corbett, A. T., Gowda, S. M., Wagner, A. Z., MacLaren, B. M., Kauffman, L. R., Mitchell, A. P. & Giguere, S. (2010). Contextual slip and prediction of student performance after use of an intelligent tutor. *Proceedings of the 18th Annual Conference on User Modeling,* an intelligent tutor. *Adaptation, and Personalization*, 52–63.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off–task behavior in the cognitive tutor classroom: When students "game the system". *Proceedings of ACM CHI 2004: Computer–Human Interaction*, 383–390.

Baker, R. S. J. d., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User–Adapted Interaction, 18*(3), 287–314.

Baker, R. S. J. d., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment–by–moment. *International Journal of Artificial Intelligence in Education, 21*(1–2), 5–25.

Baker, R. S. J. d., Gowda, S., & Corbett, A. T. (2011a). Towards predicting future transfer of learning. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 23–30.

Baker, R. S. J. d., Gowda, S. M., & Corbett, A. T. (2011b). Automatically detecting a student's preparation for future learning: Help use is key. *Proceedings of the 4th International Conference on Educational Data Mining*, 179–188.

Baker, R. S. J. d., Hershkovitz, A., Rossi, L. M., Goldstein, A. B., & Gowda, S. M. (2013). Predicting robust learning with the visual form of the moment–by–moment learning curve. *Journal of the Learning Sciences, 22*(4), 639–666.

Baker, R., & Siemens, G. (in press). Educational data mining and learning analytics. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences (2nd ed.).*

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long–term learning. *Assessment & Evaluation in Higher Education, 31*(4), 399–413.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A Iran–Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 22, pp. 61–100). Washington, DC: American Educational Research Association.

Buckingham Shum, S., & Deakin Crick, R. (2012). Learning dispositions and transferable competencies: Pedagogy, modelling and learning analytics. *Proceedings of the 2nd International Conference on Learning Analytics & Knowledge*. New York, NY: ACM.

Chin, D. B., Dohmen, L. M., Cheng, B. H., Oppezzo, M. A., Chase, C. C., & Schwartz, D. L. (2010). Preparing students for future learning with teachable agents. *Educational Technology Research and Development, 58*(6), 649–669.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User–Adapted Interaction, 4*, 253–278.

Corbett, A., Kauffman, L., MacLaren, B., Wagner, A., & Jones, E. (2010). A cognitive tutor for genetics problem solving: Learning gains and student modeling. *Journal of Educational Computing Research, 42*(2), 219–239.

Corbett, A., MacLaren, B., Wagner, A., Kauffman, L., Mitchell, A., & Baker, R. (2013a). Enhancing robust learning through problem solving in the genetics cognitive tutor. Poster paper. *Proceedings of the Annual Meeting*

of the Cognitive Science Society, 2094–2099.

Corbett, A., MacLaren, B., Wagner, A., Kauffman, L., Mitchell, A., Baker, R. S. J. d. (2013b). Differential impact of learning activities designed to support robust learning in the genetics cognitive tutor. *Proceedings of the 16th International Conference on Artificial Intelligence and Education*, 319–328.

Feng, M., Heffernan, N. T., & Koedinger, K. R. (2009). Addressing the assessment challenge in an online system that tutors as it assesses. *User Modeling and User–Adapted Interaction: The Journal of Personalization Research, 19*(3), 243–266.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120*, 34–45.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–28.

Hanna, D. E. (1998). Higher education in an era of digital competition: Emerging organizational models. *Journal of Asynchronous Learning Networks*. http://www.aln.org/alnweb/journal/vol2_issue1/hanna.htm

Hershkovitz, A., Baker, R. S. J. d., Gowda, S. M., & Corbett, A. T. (2013). Predicting future learning better using quantitative analysis of moment–by–moment learning. *Proceedings of the 6th International Conference on Educational Data Mining*, 74–81.

Jastrzembski, T. S., Gluck, K. A., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. *Proceedings of the 2006 Interservice/Industry Training, Simulation, and Education Conference*, 1498–1508.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Kinnebrew, J. S., & Biswas, G. (2012). Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution. *Proceedings of the 5th International Conference on Educational Data Mining*. Chania, Greece.

Knapper, C. K., & Cropley, A. J. (2000). *Lifelong learning in higher education* (3rd ed.). London, UK: Kogan Page.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–78). Cambridge, UK: Cambridge University Press.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge–learning–instruction (KLI) framework bridging the science–practice chasm to enhance robust student learning. *Cognitive Science, 36*, 757–798.

Linn, M., & Chiu, J. (2011). Combining learning and assessment to improve science education. *Research & Practice in Assessment, 6*(2), 5–14.

Mislevy, R. J., Almond, R. G., & Lukas, J. (2004). *A brief introduction to evidence–centered design* (CSE Technical Report 632). The National Center for Research on Evaluation, Standards, Student Testing (CRESST).

Ocumpaugh, J., Baker, R. S., Kamarainen, A. M., & Metcalf, S. J. (2014). Modifying field observation methods on the fly: Metanarrative and disgust in an environmental MUVE. *Proceedings of PALE 2013: The 4th International Workshop on Personalization Approaches in Learning Environments*, 49–54.

O'Reilly, U. M., & Veeramachaneni, K. (2014). Technology for mining the big data of MOOCs. *Research & Practice in Assessment, 9*(2), 29-37.

Pardos, Z. A., Baker, R. S. J. d., Gowda, S. M., & Heffernan, N.T. (2011). The sum is greater than the parts: Enabling models of student knowledge in educational software. *SIGKDD Explorations, 13*(2), 37–44.

Puncochar, J., & Klett, M. (2013) A model for outcomes assessment of undergraduate science knowledge and inquiry processes. *Research & Practice in Assessment, 8*(2), 42–54.

Qiu, Y., Qi, Y., Lu, H., Pardos, Z. A., & Heffernan, N. T. (2011). Does time matter? Modeling the effect of time with Bayesian knowledge tracing. *Proceedings of the International Conference on Educational Data Mining (EDM)*, 139–148.

Reye, J. (2004) Student modeling based on belief networks. *International Journal of Artificial Intelligence in Education, 14*, 1–33.

Romero C., & Ventura. S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135–146.

Sao Pedro, M., Baker, R., & Gobert, J. (2012). Improving construct validity yields better models of systematic inquiry, even with less information. *Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization*, 249–260.

Sao Pedro, M. A., Gobert, J. D., & Baker, R. (2014). The impacts of automatic scaffolding on students' acquisition of data collection inquiry skills. *Roundtable presentation at American Educational Research Association*.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207–217.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129–184.

Shih, B., Koedinger, K. R., & Scheines, R. (2008). A response time model for bottom–out hints as worked examples. *Proceedings of 1st International Conference on Educational Data Mining*, 117–126.

Shute, V. J., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: MIT Press.

Singley, M. K., & Anderson, J. R. (1989). *Transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Smith, M. U. (1988, April). *Toward a unified theory of problem solving: a view from biology*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Sperling, J., & Tucker, R.W. (1997). *For–profit higher education: Developing a world–class workforce.* New Brunswick, NJ: Transaction Publishers.

Surber, J. R., & Anderson, R. C. (1975). Delay–retention effect in natural classroom settings. *Journal of Educational Psychology, 67*(2), 170–173.

VanLehn, K., Jones, R., & Chi, M. T. H. (1992). A model of the self–explanation effect. *Journal of the Learning Sciences, 2*(1), 1–59.

Vardi, M. Y. (2012). Will MOOCs destroy academia? *Communications of the ACM, 55*(11), 5.

Wang, Y., & Heffernan, N. (2011). Towards modeling forgetting and relearning in ITS: Preliminary analysis of ARRS data. *Proceedings of the 4th International Conference on Educational Data Mining*, 351–352.