Predicting Individual Differences for Learner Modeling in Intelligent Tutors from Previous Learner Activities

Michael Eagle[1], Albert Corbett[1], John Stamper[1], Bruce M. McLaren[1], Ryan Baker[3], Angela Wagner[1], Benjamin MacLaren[1], and Aaron Mitchell[2]

[1]Human-Computer Interaction Institute, [2]Department of Biological Sciences, Carnegie Mellon University
{meagle, ac21, jstamper, bmclaren, awagner, maclaren, apm1}@andrew.cmu.edu
[3]Teacher's College, Columbia University
ryanbaker@gmail.com

ABSTRACT

This study examines how accurately individual student differences in learning can be predicted from prior student learning activities. Bayesian Knowledge Tracing (BKT) predicts learner performance well and has often been employed to implement cognitive mastery. Standard BKT individualizes parameter estimates for knowledge components, but not for learners. Studies have shown that individualizing parameters for learners improves the quality of BKT fits and can lead to very different (and potentially better) practice recommendations. These studies typically derive best-fitting individualized learner parameters from learner performance in existing data logs, making the methods difficult to deploy in actual tutor use. In this work, we examine how well BKT parameters in a tutor lesson can be individualized based on learners' prior performance in reading instructional text, taking a pretest, and completing an earlier tutor lesson. We find that best-fitting individual difference estimates do not directly transfer well from one tutor lesson to another, but that predictive models incorporating variables extracted from prior reading, pretest and tutor activities perform well, when compared to a standard BKT model and a model with best-fitting individualized parameter estimates.

Keywords

BKT; Genetics; Machine Learning; Student Modeling

1. INTRODUCTION

Intelligent tutoring systems have employed learner models to improve learning outcomes for over two decades. Learner models have been used both to individualize curriculum sequencing [1, 2, 3] and/or to individualize hint messages [4, 5]. Each of the five successful modeling frameworks cited here employs a Bayesian method to infer learner knowledge from learner response accuracy, and Bayesian modeling systems have been shown to accurately predict students' tutor and/or posttest performance [1, 3, 6, 7].

Bayesian models generally individualize model parameters for different reasoning skills, or *knowledge components*, (KCs), but not for different students. Several studies have shown that individualizing parameters for students, as well as for KCs, improves the quality of the models [1, 8, 9, 10].

These modeling studies of individual differences among students have employed data sets consisting of tens of KCs, or even many hundreds of KCs [10]. These studies have analyzed students' performance on a set of KCs retroactively, deriving the individualized student parameters for that set of KCs from existing tutor log files. These methods successfully address the research question, but are complicated to use for actual student modeling in ITSs, since the concurrent estimation and use of individualized parameters within a tutor lesson can be quite challenging at best.

In this paper we examine whether parameter estimates can be individualized for students prior to embarking on a tutor module, based on student performance in earlier activities. First, we examine whether parameter estimates can be individualized based on performance in two activities that naturally precede tutor use: reading on-line instructional text and taking a conceptual knowledge pretest.

Second, we examine whether, once a student begins using an ITS, parameter estimates in a prior tutor module can be individualized based on student performance in a prior module. In particular, how well do individualized student parameters directly transfer from one tutor module to the next? If not well, what measures of student performance in a tutor lesson can be used to predict individual student parameters in a following lesson?

We explore this issue in the Bayesian Knowledge Tracing modeling framework [1] and in a unit of the Genetics Cognitive Tutor [6]. In the following sections we describe Knowledge Tracing, the on-line student activities, and the predictors derived from students' reading, pretest, and prior tutor activities. Finally, we report our success in using these predictors to model individual differences in student learning and performance in the tutor.

2. MODELING FRAMEWORK

Bayesian Knowledge Tracing (BKT) [1] employs a two-state Bayesian learning model for each knowledge component (KC) in a tutor curriculum: at any time a student either has learned or not learned a given KC. BKT employs four parameters to estimate the probability that a student has learned each KC:

- pL_0 *initial knowledge* the probability a student has learned how to apply a KC prior to the first opportunity to apply it in the ITS
- pT *learning rate* the probability a student learns a KC at each opportunity to apply it
- pG *guessing* the probability a student will guess correctly if the KC is not learned
- pS *slips* the probability a student will make an error when the KC has been learned

Cognitive Tutors employ BKT to implement cognitive mastery, in which the curriculum is individualized to afford each student just the number of practice opportunities needed to enable the student to "master" each of the KCs. Mastery is generally operationalized as a 0.95 probability that the student has learned the KC.

UMAP'16, July 13–17, 2016, Halifax, Nova Scotia, Canada.

DOI: http://dx.doi.org/10.1145/2930238.2930255

2.1 Individual Differences

Knowledge Tracing generally employs best-fitting estimates of each of the four parameters for each KC but not for individual students. In this work, we incorporate individual differences among students into the model in the form of individual difference weights. Following Corbett & Anderson [1], four best-fitting weights are estimated for each student, one weight for each of the four parameter types, wL_0, wT, wG, wS. In estimating and employing these *individual difference weights* (IDWs), each of the four probability estimates for each rule is converted to odds form (p/(1-p)), multiplied by the corresponding student-specific weight and the resulting odds form is converted back to a probability. Let i represent the parameter type, (i.e., pL_0, pT, pG, pS), r represent the reasoning rule (KC) and s the student, then the individually weighted parameter for each rule and student, $p_{irs,}$ is given by the equation:

$$p_{irs} = p_{ir} * w_{is} / (p_{ir} * w_{is} + (1 - p_{ir})) \quad (1)$$

where p_{ir} is a best fitting parameter estimate for the rule across all students and w_{is} is the corresponding individual difference weight for the students.

2.2 Related Work

Several previous studies have employed tutor log files to retroactively examine the impact of individualizing BKT parameters for students. Corbett and Anderson [1] individualized all four BKT parameters for students, as described in Section 2.2, and found that the resulting model predicted individual differences in post-test performance better than the standard, non-individualized BKT model. Lee & Brunskill [8] employed a different method to derive four individual difference parameters and examined the impact on another property of the models – the number of practice opportunities that would be required to reach mastery. They found that the individualized model recommended substantially greater practice for some students and substantially less practice for others than the standard, non-individualized model. Two other studies focused on individualizing just the learning parameters, pL_0 and pT and obtained somewhat different results. Pardos and Heffernan [9], individualized the *initial knowledge* parameter, $pL_{0,}$ alone, based on either the student's first attempt at each KC or on all attempts at each KC – and found that either individualized method yielded reliably better fits than the standard, non-individualized BKT model. Finally, Yudelson, Koedinger and Gordon [10] individualized both the learning parameters, and found that individualizing the *learning rate* parameter pT yielded reliably better fits than the standard, non-individualized BKT model. However, unlike Pardos and Heffernan [9], they found that individualizing pL_0, alone or along with pT, did not reliably improve the goodness of fit. Finally, in an alternative approach to modeling student differences, a variety of student modeling frameworks grounded in Item-Response Theory have employed a single individual difference parameter as a basic component of the model [11,12,13].

All of these studies model student differences with log files after the students have completed the tutor activities. In this paper, we examine whether individual student differences can be estimated before students start using a tutor lesson, based on student performance in prior activities that are natural components of on-line learning activities.

3. GENETICS TUTOR

This study employed the Genetics Cognitive Tutor [6]. This tutor consists of over 25 lessons that support problem solving across a wide range of topics in genetics, including Mendelian transmission, pedigree analysis, gene mapping, population genetics and genetic pathways analysis. Various subsets of the modules have been piloted at 15 universities and four high schools in the U.S.

The genetics topic in this study is gene interaction, which examines how two genes can interact in controlling a single phenotypic trait (an observable trait, e.g., hair color). When two genes, each with a dominant and recessive allele, control a single trait, e.g., bell pepper color, there can be up to four different resulting phenotypes (four colors). However, there are many ways the two genes can interact that result in only two or three different phenotypes.

The study employs two gene interaction lessons, which require that students reason about the topic in two different ways. In the first, forward reasoning or *process modeling* lesson, each problem provides a description of how two genes interact and students determine the phenotype that is associated with each genotype and the offspring phenotype rates that will result from various parental crosses. In the second, *abductive reasoning* lesson, students analyze offspring phenotype rates that result from various parental crosses, and reason backwards to infer the genotypes of the parents and the offspring, and ultimately, how the two genotypes interact to determine phenotype.

This study focuses on four on-line activities that students completed in succession: reading the gene interaction instructional text online, taking a gene interaction pretest, and finally using both the Genetics Cognitive Tutor Gene Interaction process modeling and abductive reasoning modules.

On-Line Instructional Text. The online instructional text consisted of 23 screens, structured like pages in a book. Students could move forward and backward through the text, one screen at a time. After a student touched each page at least once, a "done" button appeared on the final (23rd) screen and the student could then continue reading (e.g., back up to re-read pages), or exit at any time.

Conceptual Knowledge Pretest. Students completed a pretest with nine conceptual questions divided into three topics. The first three questions focused on general knowledge of basic Mendelian transmission with 2 genes, the second three questions focused on process modeling, and the last three questions focused on abductive (backward) reasoning. This was not a problem-solving pretest; the last six questions are not similar to the Cognitive Tutor problems. Instead, they required students to reason about genetics processes and abductive reasoning more abstractly.

Genetics Cognitive Tutor: Gene Interaction Process Modeling. This lesson consisted of 5 problems. In each problem, students were given a description of how two genes interact to determine a phenotype, e.g., bell pepper color. Students (a) mapped the description onto one of seven gene interaction templates with 3 menus, (b) identified the phenotypes of the four true-breeding genotypes. (c) modeled the offspring genotypes and phenotypes resulting from two different parental crosses, and finally (e) summarized the phenotypes associated with all possible individual genotypes and how the phenotypes arise.

Genetics Cognitive Tutor: Gene Interaction Abductive Reasoning. This lesson consisted of 6 problems. Each problem, displayed the offspring phenotype rates that result from 3 parental crosses. Students inferred the genotypes of the parents and offspring in each of the three crosses and how the two genes interact to determine genotypes

The Cognitive Model for the Two Lessons. There was an average of 45 steps in each of the process modeling problems and 25 steps in each of the abduction problems. Some of the KCs governing the steps in a problem were unique to the problem, while others were applicable in multiple problems. In this analysis we excluded KCs that occurred only once or twice across the problems in a lesson, leaving 31 KCs in the process modeling lesson and 22 KCs in the abduction lesson.

3.1 Predictor Variables

In this study, we examine the effectiveness of four categories of student performance variables in predicting Lesson 2 (abduction) IDWs: (1) reading the instructional text, (2) pretest performance, (3) Lesson 1 IDWs and (4) features of student performance in completing Lesson 1.

In a prior paper [14] we derived 12 predictor variables for the students in this in this study, based on their gene interaction reading and pretest activities – 6 reading variables and 6 pretest variables, as described in this section. In that paper these 12 measures were used to predict best fitting IDWs for tutor Lesson 1 (process modeling), as summarized in section 3.2 below.

3.1.1 Predictors Derived from Instructional Text and Reading Performance

We derived two types of measures of student reading performance: reading time and revisiting pages in the text. Between these two measures we derived a total of 6 predictor variables, as follows.

Reading Time (4 variables): No prior ITS research employs reading rates to individualize parameters in a learning environment, but there is substantial evidence that reading time may prove sensitive to individual differences in comprehension difficulty. Harvey and Anderson [15] showed that reading times for on-line declarative instruction in the ACT Programming Tutor are sensitive to differences in processing time necessary to encode familiar versus novel material. More generally, an extensive research literature demonstrates that, reading time is sensitive to relative comprehension difficulty [16].

We performed a factor analysis on log reading times for the 23 individual pages to reduce the number of predictors. The factor analysis yielded a total of four factors (see RTF1, RTF2, RTF3, RTF4 in Table 1), which align with subtopics in the text, as summarized in Table 1.

Text Pages Revisited (2 variables): Students can read through the declarative instruction as they would pages in a book. Some students may choose to strictly read forward through the text, while others may choose to revisit earlier pages in the text. We calculated two measures of student behavior in revisiting text pages: the number of pages re-read and number of intervening pages traversed in re-reading text pages.

3.1.2 Predictors Derived from a Conceptual Knowledge Pretest

Some prior projects have employed pretest accuracy to initialize ITS student models [3, 17]. We derived three types of measures of student pretest performance: accuracy, answer changes and time on task. Between these three measures, we derived a total of 6 predictor variables, as follows.

Pretest Accuracy (3 variables): We calculated students' average pretest accuracy on each of the three types of pretest questions, general knowledge, process modeling and abductive reasoning.

Pretest Answer Changes (2 variables): We calculated the number of times students changed their answers in the pretest from a correct initial answer to an incorrect final answer, or vice versa.

Time on Task (1 variable): Finally we calculated students' total time to complete the pretest.

3.1.3 Lesson 1 Individual Difference Weights

In the prior study [14] we derived four best-fitting individual difference weights in the first tutor lesson for each of the students in this study. In this study we examine both how well these Lesson 1 IDWs directly apply to Lesson 2, and whether they can be used to improve the predictive model for Lesson 2 IDWs.

3.1.4 Lesson 1 Performance Features

Finally, we derived six features of student performance in solving the Lesson 1 tutor problems.

Error Rate: Corbett and Anderson [1] found that students' raw error rate within a tutor lesson is strongly correlated with the logarithm of students' IDWs for that lesson. In this study we calculated students' raw error rate in completing the Lesson 1 problems and examine whether Lesson 1 error rate predicts Lesson 2 IDWs.

Average response time: We calculated students' average response time for their first problem-solving action at each opportunity to apply one of the 22 KCs in tutor Lesson 2.

3.1.4.1 Performance Features that Predict Transfer and Preparation for Future Learning

Predicting individual differences in students' initial knowledge, pL_0, in a tutor lesson from their performance in a prior tutor lesson is closely related to examining the direct *transfer* of knowledge from the first lesson to the second lesson. Similarly, predicting individual differences in students' learning rate, pT, in a lesson from their performance in a prior tutor lesson is closely related to examining students' *preparation for future learning* [18] after completing the first lesson.

Prior research [19] has identified features of students' performance in a Genetics Cognitive Tutor lesson that predict transfer and preparation for future learning, which are both manifestations of deep or "robust" student understanding [20]. Conversely, features that predict shallow learning have also been identified [21]. In this study we examine four performance features that correlate with at least two of these three constructs.

Help Avoidance: The proportion of problem-solving steps in which the probability that the student knows the relevant KC is low and the first action is an error instead of a hint request.

Bug Message Long Pause: The proportion of a student's actions in which a bug message (error messages given when the student's behavior indicates a known misconception) is followed by a long pause before a subsequent action.

Hint Long Pause: The proportion of a student's actions in which a hint request is followed by a long pause before the next action.

Hint Correct Long Pause: The proportion of a student's actions in which a hint request is followed by a correct action and then a long pause before the next action.

3.2 Prior Gene Interaction Lesson 1 Results

Eagle, et al, [14] examined the feasibility of setting individual difference weights for the first lesson in this tutor curriculum sequence, before students begin work in the lesson. That study derived best-fitting standard BKT parameters for each of the 31 KCs in the first gene interaction lesson, as described in Section 2

above, calculated best-fitting IDWs for each of the students, as described in Section 2.1, then examined how well each student's best-fitting IDWs could be predicted from the 12 reading and pretest variables described in sections 3.1.1 and 3.1.2.

The BKT model with *best-fitting* IDWs (FIDW-31) improved the goodness of fit of the standard non-individualized BKT model by 8.7%, reducing the RMSE from 0.306 to 0.279. The BKT model with *predicted* IDWs (PIDW-31) was about 40% as successful as the best-fitting model. It improved the goodness of fit of the standard non-individualized BKT model by 3.6%, reducing RMSE from 0.306 to 0.295.

As Koedinger, et al [22] observed, even small differences in model fits can have large effects on the amount of recommended work assigned to the student. To compare the practical impact of the individualized best-fitting FIDW-31 and predicted PIDW-31 models for Lesson 1, we calculated the number of practice opportunities that would be needed for students to reach mastery under the standard non-individualized BKT model (SBKT), and under the FIDW-31 and PIDW-31 models – that is the number of opportunities that would be required for *pL* (the probability the students has learned a KC) to reach the mastery criterion (0.95). While students completed a fixed curriculum in this study, most students had in fact reached the mastery criterion for most of the KCs.

Under the FIDW-31 individualized model, 56 students required less practice to reach mastery than under the SBKT non-individualized model and these students required 17 fewer opportunities on average. Under the FIDW-31 model, 27 students needed more practice than under the SBKT model to reach mastery. These students required an average of 27 fewer opportunities.

There was substantial, but not perfect agreement between the predicted PIDW-31 and best-fitting FIDW-31 models on the amount of practice individual students needed to achieve mastery. The PIDW-31 model recommended less practice then the SBKT model for 54 students, vs. 56 for FIDW-31, and the two models agreed on 46 of these students. The PIDW-31 model recommended less practice to reach mastery for 27 students, and the two models agreed on 19 of these students. However, the PIDW-31 only recommended 11 fewer opportunities for the first group (vs. 17 for the FIDW-31 model) and only 14 more opportunities for the latter group (vs. 27 for the PIDW-31 model).

Given this moderate level of success in predicting IDWs from prior activities before students embark on the first tutor lesson, in the current study we examine two questions. (1) Should IDWs be estimated separately for successive lessons in a tutor curriculum? (2) Can we predict IDWs in the second lesson more accurately if we employ predictor variables from student performance in the first tutor lesson, as well as from reading and pretest activities?

4. METHODS AND MATERIALS

The data analyzed in this study come from 83 CMU undergraduates enrolled in either genetics or introductory biology courses who were recruited to participate in this study for pay. Students participated in two 2.5-hour sessions on consecutive days in a campus computer lab. In this study, the first session focused on gene interaction and students read the on-line gene interaction instructional text, took the on-line pretest, and used the gene interaction process modeling tutor module and the abductive reasoning tutor module as the first four activities in this session. The study focuses on modeling the 83 students' first actions on 10,309 problem-solving steps in the abduction module.

4.1 Fitting Procedures

We first found best-fitting group parameter estimates for each of the 4 parameters (pL_0, pT, pG, pS) in the standard BKT model for each of the 22 KCs in Lesson 2, with nonlinear optimization. The objective function takes the observed opportunities for a single skill and a set of group parameters as input and returns the negative log-likelihood (-LogLik). Optimization ultimately returns the set of group parameters that best fit the skill. Both pG and pS were bounded to be less than 0.5, as in [23] to avoid paradoxical results that arise when these performance parameters exceed 0.5 (e.g., a student with a higher probability of knowing a KC is less likely to apply it correctly.)

Second, we re-fit the lesson 2 tutor data with an individualized BKT model: We obtained four best-fitting Individual Difference Weights (IDWs) for each of the 83 students, one weight for each of the four parameter types, *wL0*, *wT*, *wG*, *wS*. As described in Section 2.1 equation 1, each student's four weights are mapped across the best-fitting group learning and performance parameter estimates for each of the 22 KCs in the lesson to individualize these parameter estimates. The objective function takes the fixed group parameters, the observed opportunities for a student, and a set of IDWs (*wL0*, *wT*, *wG*, *wS*) and returns the -LogLik. Optimization ultimately returns the set of IDWs that maximize the fit.

Table 1. 22 Predictor variables employed in this study.

Reading Predictors (from Eagle, et al [14])

1. **RTF1** Reading: Time for a 5-page intro with familiar content on basic Mendelian genetics
2. **RTF2** Reading: Time for 6 pages with charts of various ways 2 genes can interact
3. **RTF3** Reading: Time for 3 pages on parental crosses with offspring genotypes & traits
4. **RTF4** Reading: Time for 2 pages with full-page diagrams of dominant & recessive alleles
5. **RRNP** Reading: Total number of previous pages re-read
6. **RRTD** Reading: Total distance traversed (intervening pages) in re-reading text pages

Pretest Predictors (from Eagle, et al [14])

1. **PACC1** Pretest: % Correct for 3 general knowledge questions
2. **PACC2** Pretest: % Correct for 3 process modeling questions
3. **PACC3** Pretest: % Correct for 3 abductive reasoning questions
4. **PCCI** Pretest: Number of answers initially correct changed to incorrect
5. **PCIC** Pretest: Number of answers initially incorrect changed to correct
6. **Ptime** Pretest: Total time to complete the pretest

Tutor Lesson 1 Individual Difference Weights

1. **L1wL_0** Lesson 1 Initial Learning IDWs
2. **L1wT** Lesson 1 Learning Rate IDWs
3. **L1wG** Lesson 1 Guessing IDWs
4. **L1wS** Lesson 1 Slip IDWs

Tutor Lesson 1 Predictors

1. **TErr** Lesson 1 proportion of errors
2. **TTime** Lesson 1 average response time
3. **HELPA** Not requesting help on poorly learned skills
4. **BugLP** Bug message followed by a long pause
5. **HNLP** Hint message followed by a long pause
6. **HNLPC** Hint message followed by a correct action then a long pause

Third, we derived 6 features from student performance in Lesson 1, as described in section 3.1.4. Along with the 12 reading and pretest features and 4 best-fitting individual difference weights derived in [14], this yields a total set of 22 predictor variables, displayed in Table 1.

Fourth, we employed these variables to independently predict the four Lesson 2 IDWs: wL_0, wT, wG, wS. We generated three predictions for each of the IDWs with successively larger subsets of features: (1) the 12 reading and pretest features; (2) the 12 reading and pretest features and 4 Lesson 1 IDWs; (3) the 12 reading and pretest features, the 4 Lesson 1 IDWs, and the 6 Lesson 1 performance features.Since we are predicting multiplicative weights, we fit a transformation of the weights $w/(1+w)$. This transformation has the property that the neutral weight 1.0 (which does not modify the corresponding best-fitting group parameter), is the midpoint of the transformed scale.

4.2 Model and Feature Selection

In order to reduce the number of features, and to compare to the previous work in [14] we used Least Angle Regression (LAR) [24] a variant of Lasso. For each of the four Lesson 2 IDWs we use LAR to select the best 12 predictors (out of 22.) Lasso performs both variable selection and regularization, and restricts the size of the coefficients making some of the values be zero (not included in the model.)

We then built a robust regression model with the 12 predictors for each of the IDWs. Robust regression is less sensitive to outliers, variable normality, and other violations of standard linear regression assumptions [25].

Finally, we employed the various sets of predictors to calculate 5 new IDW BKT models, yielding a total of six BKT model variants displayed in Table 2. Analysis work was performed using R [26], Optimx [27], rlm [28], and lars [24].

Table 2. Six Lesson 2 BKT models calculated in this analysis

1. **SBKT**: Standard BKT non-individualized model with best-fitting group parameter estimates.
2. **FIDW-22** Individualized BKT model with Fitted Individualized Difference Weights from Lesson 2
3. **FIDW-31**: Individualized BKT model with Fitted Individualized Difference Weights from Lesson 1
4. **PIDW-RP**: Individualized BKT model with predicted IDWs from reading and pretest features.
5. **PIDW-RPW**: Individualized BKT model with IDWs predicted from 12 reading and pretest features and 4 Lesson 1 IDWs
6. **PIDW-RPWF**: Individualized BKT model with IDWs predicted from 12 reading and pretest features, 4 Lesson 1 IDWs and 6 Lesson 1 performance features.

5. RESULTS

This section examines three main questions:

- How well do best-fitting IDWs transfer from one tutor lesson to a following tutor lesson?
- Do features of reading and pretest performance still predict best-fitting IDWs in a tutor lesson following an intervening tutor lesson?
- Do performance features from a prior tutor lesson further improve predicted IDWs in a subsequent tutor lesson?

5.1 Best-Fitting Models and Generalizability of Individual Difference Weights

Table 3 displays the overall fit to student performance in tutor Lesson 2 of three best-fitting BKT models. Column 2 displays root mean squared error (RMSE) and column 3 displays accuracy (the probability a model correctly predicts whether a student response will be correct or incorrect, with a 0.5 threshold on predicted accuracy.) The first row in the table displays the standard BKT model (SBKT) with no individualization for students as a baseline.

Table 3. Goodness of fit of 3 models for Lesson 2 tutor data: The standard BKT model & 2 BKT models with lesson-specific IDWs

Model	RMSE	Accuracy
SBKT 22 KCs	0.413	0.749
Lesson 2 FIDW-22 KCs	0.385	0.784
Lesson 1 FIDW-31 KCs	0.415	0.756

The last two rows in Table 3 display the goodness of fit of two BKT models that incorporate best-fitting IDWs. The second row displays the BKT model with IDWs trained on the 22 KCs in Lesson 2 (FIDW-22). As can be seen, this model improves the goodness of fit compared to the SBKT model, reducing RMSE by about 6.8% (RMSE 0.385 vs. 0.413) and increasing accuracy by about 4.7% (Accuracy 0.784 vs. 0.749).

The last row examines the generalizability of IDWs across the two tutor lessons. This row displays the BKT model for lesson 2 with the IDWs that were previously trained on the 31 IDWs in lesson 1 in [14]. As can be seen, the IDWs trained on lesson 1 KCs do not transfer well to lesson 2. The overall RMSE for this individualized model is slightly worse than for the non-individualized SBKT model, while the Accuracy is somewhat better. Even for these two highly related tutor lessons, which require students to reason differently about the same genetics knowledge, simply propagating IDWs from one lesson to another is not successful.

5.2 Predicting Lesson 2 IDWs

In this section we evaluate three methods for predicting Lesson 2 IDWs from student performance with three activities that precede Lesson 2: reading instructional text, a pretest, and tutor Lesson 1. We employ the FIDW-22 model with best-fitting lesson-specific IDWs as our gold standard for evaluating model fits with predicted IDWs. Table 4 displays the overall goodness of fit of these three models. (The first two rows display the fit of the standard non-individualized SBKT model, and best fitting BKT model with lesson-specific IDWS, FIDW-22, for comparison.)

Table 4. Goodness of fit of 5 models for lesson 2 tutor data: The standard BKT model, 4 BKT models with lesson-specific IDWs.

Model	RMSE	Accuracy
Lesson 2 SBKT	0.413	0.749
Lesson 2 FIDW-22	0.385	0.784
Lesson 2 PIDW-RP	0.399	0.764
Lesson 2 PIDW-RPW	0.397	0.769
Lesson 2 PIDW-RPWF	0.396	0.769

5.2.1 Predicting Lesson 2 IDWs with Reading and Pretest Variables

We first employed the six reading measures and six pretest measures derived previously [14] to predict Lesson-2 specific IDWs, as described in Section 4.1. The overall goodness of fit for this model PIDW-RP is displayed in row 3 of Table 4. This model with predicted Lesson 2 IDWs is 50% as successful as the best-fitting FIDW-22 model in reducing RMSE: The new model reduces RMSE by 3.4% compared to the non-individualized SBKT model, 0.399 vs. 0.413, vs. a 6.8% improvement for the best-fitting FIDW-22 model). The FIDW model is also about 2.0% more accurate than the SBKT model, 0.764 vs. 0.749, (vs. a 4.7% improvement for the best-fitting FIDW-22 model).

Table 5. Differences in practice needed to reach mastery.

Model	# Stus. needing less	# Fewer Opps. Needed	# Stus. needing more	# More Opps. Needed
FIDW-22	49	16.69	33	12.21
PIDW-RP	57 (40)	4.73	22 (14)	7.21
PIDW-RPW	54 (38)	4.35	23 (15)	8.28
PIDW-RPWF	56 (40)	4.65	29 (16)	8.96

As discussed earlier, small differences in model fits can have large effects on the amount of recommended work assigned to the student [22]. To compare the practical impact of the best-fitting FIDW-22 model and the three predicted IDWs, we calculated the number of practice opportunities that were necessary for students to reach mastery under each of the models - that is, the number of opportunities required for *pL* (the probability the student has learned a rule) to reach 0.95. While students completed a fixed curriculum in this lesson, this analysis is possible because most students reached mastery for most of the KCS in the available number of opportunities under all three models. Across all students and skills, students mastered 75% of the skills under the SBKT model, 77% under the FIDW model, and 77% under the PIDW-RP model. If a student did not reach mastery on a KC under one model, we conservatively estimated that the student would reach mastery on the next opportunity. This means that the number of *More Opps* is a lower bound and interpreted as the minimum number of opportunities the model would recommend.

The practice recommendations are displayed in Table 5. The second column displays how many students would need less practice under the individualized model than under the non-individualized SBKT model. The third column displays how many fewer practice opportunities these students would need on average. The fourth column displays how many students would need more practice under the individualized model than under the non-individualized SBKT model. The fifth column displays how many more opportunities would be needed on average.

Both the best-fitting FIDW-22 model in row 2 and the predicted PIDW-RP model in row 3 substantially modify the amount of practice students need to reach mastery compared to the SBKT model. Under the best-fitting FIDW model, 49 students needed less practice to master all the KCs than under the non-individualized SBKT model and on average these students required 16.69 fewer practice opportunities to reach mastery under FIDW than under SBKT. Under the predicted PIDW-RP model, 57 students needed an average of 4.73 fewer opportunities to master all the KCS than under the SBKT model. The two individualized model agree on a set of 40 students who need fewer practice opportunities to reach mastery, but again the FIDW model requires less practice (16.69 opportunities) of these students than the PIDW-RP model (4.73 opportunities).

Under the best-fitting FIDW model, 33 students needed more practice to master all the KCs than under the non-individualized SBKT model and on average these students required 12.21 fewer practice opportunities to reach mastery under FIDW than under SBKT. Under the predicted PIDW-RP model, 22 students needed an average of 7.21 fewer opportunities to master all the KCS than under the SBKT model. The two individualized model agree on a set of 14 students who need fewer practice opportunities to reach mastery, but again the FIDW model requires less practice (18.38 opportunities) of these students than the PIDW-RP model (7.58 opportunities).

Overall, the FIDW and PIDW-RP models were in 65% agreement on which students needed fewer or more opportunities to master all the KCs than under the SBKT model, but the new predicted PIDW-RP model is not realizing all the learning efficiency gains identified by the best-fitting FIDW model.

5.2.1.1 Models with Reading and Pretest Variables.

Table 6 displays the coefficients for each of the 12 predictors in the regression models for each of the four Lesson 2 IDWs. The predictors that enter reliably into the robust regression models are highlighted with asterisks.

Both reading time variables and pretest variables continue to be reliable predictors of individual difference weights in gene interaction tutor lesson 2, even after students have completed an intervening tutor lesson. All four reading time factors each reliably predicted at least one of the four individual differences weights. Both variables that measure the extent to which students revisit text pages also marginally predict *wG*.

Not surprisingly, pretest accuracy variables reliably entered into the four IDW models. Differences in student accuracy on general knowledge (PACC1) and on process-modeling (PACC2) each reliably predict three of the four IDWs. Surprisingly, student accuracy on abductive reasoning questions (PACC3), the type of reasoning employed in this second tutor lesson did not reliably predict any lesson 2 IDWs. Pretest reasoning about process modeling is a better predictor of students acquiring abductive reasoning skills than a pretest measure of abductive reasoning. Finally, the number of answer changes students made and total time did not reliably predict any of the four IDWs.

5.2.2 Predicting Lesson 2 IDWs from Student Reading, Pretests and Lesson 1 IDWs

While the best-fitting Lesson 1 IDWs fit the Lesson 2 data poorly, our next predictive model includes them along with the 12 reading and pretest variables. The overall goodness of fit for this model PIDW-RPW is displayed in row 4 of Table 4. This PIDW-RPW model improves upon the predictive accuracy of the earlier PIDW-RP model. The PIDW-RPW model is 57% as successful as the best-fitting FIDW-22 model both in reducing RMSE (vs. 50% for the earlier PIDW-RP model) and in increasing Accuracy (vs. 43% for the earlier PIDW-RP model). The new model reduces RMSE by 3.9% compared to the non-individualized SBKT model, (0.397 vs. 0.413) and increases accuracy by 2.7% compared to the SBKT model (0.769 vs. 0.749).

The practice recommendations for this PIDW-RPW model are displayed in row 3 of Table 5. Like the earlier individualized BKT fits, this fit substantially modifies the amount of practice students need to reach mastery compared to the SBKT model. While this new PIDW-RPW predictive model fits the tutor data better than the earlier PIDW-RP model, the practice recommendations of the two predictive models are very similar. Overall, the best-fitting FIDW and PIDW-RPW models were in 64% agreement on which students needed fewer or more opportunities to master all the KCs than under the non-individualized SBKT model, but the new predicted PIDW-RP model is not realizing all the learning efficiency gains identified by the best-fitting FIDW model.

Table 6. Coefficient Summary Table (*<0.10, **<0.05, ***<0.01)

	wL0	wT	wG	wS
(Intercept)	0.501***	0.559***	0.511***	0.509***
RTF1	-0.027	-0.047	-0.013	0.037***
RTF2	-0.016	0.076**	-0.005	-0.048***
RTF3	-0.052**	-0.025	0.023	-0.017
RTF4	0.04*	-0.017	0.03	-0.013
RRTD	-0.021	0.072	-0.105*	-0.025
RRNP	0.032	-0.088	0.115*	0.039
PACC1	0.06**	0.012	0.074**	-0.054***
PACC2	0.093***	0.084**	0.051	-0.046***
PACC3	0.017	-0.015	-0.002	0.005
PCCI	-0.001	-0.039	-0.046	0.006
PCIC	-0.032	0.022	-0.005	0.003
Ptime	-0.012	0.001	-0.008	0.021
RMSE	0.186	0.217	0.233	0.109

5.2.2.1 The Predictive Models with Reading and Pretest Variables and Lesson 1 IDWs.

Table 7 displays the coefficients for each of the 16 predictors in the regression models for each of the four Lesson 2 IDWs. As described in Section 4.1, Lasso was used to identify the best 12 predictors for each of the four IDWs. The predictors that enter reliably into the four robust regression models are highlighted with asterisks. All four Lesson 1 IDWs enter into at least two of the models. Each of the two learning weights, L1wL_0, and L1wT reliably predicts the corresponding weight in Lesson 2, but the two Lesson 1 performance weights do not reliably predict the performance weights in Lesson 2. The pattern of reading predictor variables that reliably predict each of the four weights is very similar between the first and second predictive models. But with the inclusion of the Lesson 1 IDWs, the first two pretest variables no longer enter reliably into the predictive models for the two learning weights, wL_0, and wT, nor into the guessing weight, wG, although they continue to predict the slip weight, wG, reliably. Thus, overall, information on individual differences in students' Lesson 1 learning and performance largely replaces pretest assessments of student knowledge in predicting Lesson 2 IDWs.

5.2.3 Predicting Lesson 2 IDWs from Reading, Pretests, Lesson 1 IDWs, and performance features

Our final model examines whether predictive model accuracy is further improved by including the six features of student performance in Lesson 1, along with the reading and pretest features and Lesson 1 IDWs. The overall goodness of fit for this model PIDW-RPWF is displayed in row 5 of Table 4. This PIDW-RPWF does not markedly improve the overall goodness of fit, compared to the prior PIDW-RPW model. This PIDW-RPWF is 60% as successful as the best-fitting FIDW-22 model in reducing RMSE (vs. 57% for the earlier PIDW-RP model) and 57% as successful in increasing Accuracy (vs. 57% for the earlier PIDW-RP model). The new model reduces RMSE by 4.1% compared to the SBKT model, (0.396 vs. 0.413) and increases accuracy by 2.7% compared to the SBKT model (0.769 vs. 0.749).

Table 7. Coefficient Summary Table (*< 0.10, ** < 0.05, ***<0.01)

	wL0	wT	wG	wS
(Intercept)	0.497***	0.558***	0.51***	0.509***
RTF1	-0.035*	-0.036		0.021
RTF2	-0.02	0.052*		-0.043***
RTF3	-0.049**		0.009	
RTF4	0.029		0.021	
RRTD		0.075	-0.1*	
RRNP	0.011	-0.088	0.092*	0.019
PACC1		0.008	0.045	-0.038**
PACC2	0.029	0.041	0.036	-0.048**
PACC3	0.002	-0.025	-0.024	0.022
PCCI	-0.007	-0.023	-0.04	0
PCIC	-0.022			0
Ptime		-0.021		0.024*
L1wL0	0.075***		0.021	
L1wT		0.074**	0.083**	-0.035**
L1wG	0.012	0.08**	0.013	0.021
L1wS	-0.106***	0.014	-0.005	0.028
RMSE	0.153	0.205	0.222	0.100

The practice recommendations for this PIDW-RPWF model are displayed in row 4 of Table 5. Like the earlier individualized BKT fits, this model again substantially modifies the amount of practice students need to reach mastery compared to the SBKT but, again, the practice recommendations for the FIDW-RPWF model are similar to the two prior predictive models. Overall, the FIDW and PIDW-RP models were in 67% agreement on which students needed fewer or more opportunities to master all the KCs than under the SBKT model, but the new predicted PIDW-RP model is not realizing all the learning efficiency gains identified by the best-fitting FIDW model.

5.2.3.1 The Models with Reading, Pretest Variables, and Lesson 1 IDWs and performance features.

Table 8 displays the coefficients for each of the 22 predictors in the regression models for each of the four Lesson 2 IDWs. As described in Section 4.1, Lasso was used to identify the best 12 predictors for each of the four IDWs. The predictors that enter reliably into the four robust regression models are highlighted with asterisks. As can be seen, among the 22 predictors, only help avoidance did not enter into any of the four predictive IDW models, although another 8 predictors were not even marginally significant in any of the four models. With the introduction of both four lesson 1 IDW weights and 6 lesson 1 performance features as

predictive variables, five of the six reading time variables still reliably predict at least one of the four lesson 2 IDWs. However, only two of the six pretest variables enter even marginally into predicting a single lesson 2 IDW. Each of the lesson 1 IDWs reliably predicts a single lesson 2 IDW, but lesson 1 L1*wT* no longer reliably predicts three of the four lesson 2 IDWs. Among the six Lesson 1 performance variables, raw error rate and mean response time did not enter reliably into any of the four IDW models. Long pauses after bugs (and help avoidance) also did not reliably enter into any of the models. Each of the hint response-related variables entered at least marginally into predicting two of the four lesson 2 IDWs, so student responses to hints in lesson 1 provide information about individual differences in lesson 2 learning and performance over and above measures of student reading and pretest performance and students' lesson 1 IDWs.

Table 8**.** Coefficient Summary Table ($^{*} < 0.10$, $^{**} < 0.05$, $^{**}<0.01$)

	wL0	**wT**	**wG**	**wS**
(Intercept)	0.496***	0.564***	0.509***	0.508***
RTF1	-0.026	-0.031	0.018	0.015
RTF2	-0.022	0.034		-0.035**
RTF3	-0.042**		0.017	
RTF4	0.032*		0.018	
RRTD			-0.117**	
RRNP		-0.022	0.11**	
PACC1			0.03	-0.033*
PACC2	0.024	0.024	0.027	-0.041**
PACC3		-0.046		
PCCI			-0.047	
PCIC	-0.023			
Ptime				0.019
L1wL0	0.064**		-0.004	
L1wT		0.035	0.088**	-0.028
L1wG	0.011	0.069**		0.023
L1wS	-0.07**			0.022
TErr	-0.042	-0.059	-0.011	0.005
TTime		-0.04		0.01
HELPA				
BugLP	-0.018	0.015		
HNLP		0.056*	-0.077**	0.005
HNLPC	-0.033*	-0.058**		0.006
RMSE	0.146	0.198	0.214	0.101

6. CONCLUSIONS

We have examined four methods of incorporating individual student differences into a traditional Bayesian Knowledge Tracing model in an intelligent tutor lesson based on student performance in earlier on-line activities. The simplest method, of directly employing best-fitting individual difference weights (IDWs) from the preceding tutor lesson on a closely related topic, was unsuccessful. The fit of this individualized model was no better overall than the standard non-individualized BKT model.

The other three methods employed measures of student performance in reading instructional text, taking a pretest, and completing the prior tutor lesson to predict individual difference weights in the following lesson. We found that the predictive model, which only employs measures of students' performance in reading an instructional text and in taking a pretest, was quite successful. The goodness of fit of this predictive model falls midway between the non-individualized standard BKT model and the model with actual best fitting IDWs. The individualized practice recommendations for this predictive model are similar to the practice recommendations for the model with best fitting IDWs, , although this predictive model does not identify all the opportunities to decrease the amount of practice for some students, nor the need to increase the amount practice for other students, that are identified in the best-fitting model. A second predictive model which incorporates these reading and pretest variables along with the four individual weights from the prior lesson appreciably improves the goodness of fit. However, a third predictive model which includes all of these predictor variables and another six measures of student performance in the prior tutor lesson did not appreciably improve the goodness of fit of the second predictive model.

An important conclusion of this study is that student performance in reading on-line instructional text is a useful predictor of learning and performance in an intelligent tutor. In the second model, five of six reading variables entered at least marginally into the prediction of at least one IDW, even though students had completed an intervening tutor lesson and the students' IDWs from the prior lesson were incorporated into the predictive models. Not surprisingly, in the first model, several conceptual pretest variables also reliably predicted individual differences in learning and performance in the second lesson. However, when IDWs from the first lesson are incorporated into the second model, pretest measures become much less important in predicting IDWs. A final intriguing conclusion is that in the third model, students' responses to hint messages in the first lesson were a reliable predictor of individual differences in learning and performance in the second lesson.

Predicting individual student differences in a tutor lesson from prior activities is important since incorporating individual differences into a lesson is easier if they can be assigned before students starting working with the tutor. We anticipate that the quality of predicted individual differences will further increase with additional research. And, while the lesson 1 IDWs and performance features are specific to an ITS environment, we believe that reading data, as well as pretest data, could be used to predict individual difference parameters in other types of learning environments.

7. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation via the grant "Knowing What Students Know: Using Education Data Mining to Predict Robust STEM Learning", award number DRL1420609.

8. REFERENCES

1. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4, 253-278. (1995)
2. Mayo, M., Mitrovic, A. Optimising ITS behaviour with Bayesian networks and decision theory. International Journal of Artificial Intelligence in Education, 12, 124-153 (2001).

3. Shute, V.: Smart: Student Modeling Approach for Responsive Tutoring. User Modeling and User-Adapted Interaction, 5 (1), 1-44. (1995)
4. Ganeshan, R., Johnson, L., Shaw, E., Wood, B.: Tutoring diagnostic problem solving. In G. Gauthier, C. Frasson, K. VanLehn (eds.) ITS2000 Intelligent Tutoring Systems, LNCS vol. 1839, pp. 33-42. Springer, Heidelberg. (2000)
5. Conati, C., Gertner, A., VanLehn, K. Using Bayesian networks to manage uncertainty in student modeling. User Modeling and User-Adapted Interaction, 12, 371-417. (2002)
6. Corbett, A.T., MacLaren, B., Kauffman, L., Wagner, A., Jones, E. A.: Cognitive Tutor for Genetics Problem Solving: Learning Gains and Student Modeling. Journal of Educational Computing Research, 42 (2), 219-239. (2010)
7. Gong, Y., Beck, J., Heffernan, N.: Comparing knowledge tracing and performance factor analysis by using multiple model fitting. In V. Aleven, J. Kay, J. Mostow (eds.) ITS2010 Intelligent Tutoring Systems. LNCS vol. 6094, pp. 35-44. Springer, Heidelberg. (2010)
8. Lee, J., Brunskill, E.: The impact of individualizing student models on necessary practice opportunities. In: Yacef, K., Zaiane, O., Hershkovitz, A., Yudelson, M., Stamper, J. (eds.) EDM2012 Proceedings of the 5th International Conference on International Educational Data Mining Society, 118-125. (2012)
9. Pardos, Z. Heffernan, N.: Modeling individualization in a Bayesian networks implementation of Knowledge Tracing. In De Bra, P., Kobsa, A., Chin, D. (eds.) UMAP2010 Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization, LNCS vol. 6075, pp. 255-266. Springer, Heidelberg (2010)
10. Yudelson, M., Koedinger, K., Gordon, G.: Individualized Bayesian knowledge tracing models. In: Lane, C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED2013 Artificial Intelligence in Education, LNCS vol. 7926, pp. 171-180, Springer, Heidelberg. (2013)
11. Pirolli, P., Wilson, M: A theory of the measurement of knowledge content, access, and learning. Psychological Review, 105(1), 58-82. (1998)
12. Cen, H., Koedinger, K., Junker, B.: Comparing two IRT models for conjunctive skills. In B. Woolf, E. Atmour, R. Nkambou, S. Lajoie (eds.) ITS2008 Intelligent Tutoring Systems LNCS vol. 5091, pp. 796-798. Springer, Heidelberg. (2008)
13. Pavlik, P., Yudelson, M., Koedinger, K.: Using contextual factors analysis to explain transfer of least common multiple skills. In G. Biswas, S. Bull, J. Kay, A. Mitrovic (eds.) AIED2011 Artificial Intelligence in Education, LNCS vol. 6738, pp. 256-263. Springer, Heidelberg. (2011)
14. Eagle, M., Corbett, A., Stamper, J., McLaren, B.M., Wagner, A., MacLaren, B., Mitchell, A.: Estimating individual differences for student modeling in intelligent tutors from reading and pretest data. ITS2016 Intelligent Tutoring Systems. (in press)
15. Harvey, L., Anderson, J.: Transfer of declarative knowledge in complex information processing domains. Human-Computer Interaction, 11 (1), 69-96. (1996)
16. Zwann, R., Singer, M.: Text comprehension. In A. Graesser, M. Gernsbacher, S. Goldman (eds.) Handbook of discourse processes, pp. 83-121. Mahwah, NJ: Erlbaum. (2003)
17. Arroyo, I., Beck, J., Woolf, B., Beal, C., Schultz, K.: Macroadapting Animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. In G. Gauthier, C. Frasson, K. VanLehn (eds.) ITS2000 Intelligent Tutoring Systems, LNCS vol. 1839, pp. 574-583. Springer, Heidelberg. (2000)
18. Bransford, J. D., & Schwartz, D. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), Review of Research in Education (Vol. 24). Washington, DC: American Educational Research Association.
19. Baker, R.S.J.d., Corbett, A.T., Gowda, S.M. (2013). Generalizing automated detection of the robustness of student learning in an intelligent tutor for genetics. Journal of Educational Psychology. 105, 946-956.
20. Koedinger, K.R., Corbett, A.T. and Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. Cognitive Science, 36, 757-798.
21. Baker, R.S.J.d., Gowda, S.M., Corbett, A.T., Ocumpaugh, J.: Towards automatically detecting whether student learning is shallow. In S. Cerri, W. Clancey, G. Papadourakis, K. Panourgia (eds.) ITS2012 Intelligent Tutoring Systems LNCS vol. 7315, pp. 444-453. Springer, Heidelberg. (2012)
22. Koedinger, K., Stamper, J., McLaughlin, E., Nixon,T.: Using data-driven discovery of better student models to improve student learning. In: Lane, C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED2013 Artificial Intelligence in Education, LNCS vol. 7926, pp. 421-430, Springer, Heidelberg. (2013)
23. Baker, R., Corbett, A., Aleven, V.: More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In B. Woolf, E. Aimeur, R. Nkambou, S. Lajoie (eds.) ITS2008 Intelligent Tutoring Systems. LNCS vol. 5091, pp. 406-415. Springer, Heidelberg. (2008)
24. Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. Least angle regression. The Annals of statistics, 32(2), 407-499. (2004).
25. Andersen, Robert. Modern methods for robust regression. Sage, 2008.
26. Ihaka, Ross, and Robert Gentleman. "R: a language for data analysis and graphics." Journal of computational and graphical statistics 5.3, 299-314. (1996)
27. Nash, John C., and Ravi Varadhan. "Unifying optimization algorithms to aid software system users: optimx for R." Journal of Statistical Software 43.9, 1-14. (2011)
28. Venables, W. N., Ripley, B.D., Modern Applied Statistics with S. Forth Edition. Springer, New York. (2013)