Learner Differences in Hint Processing

Ilya M. Goldin
Human-Computer Interaction Institute
Carnegie Mellon University
goldin@cmu.edu

Kenneth R. Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University
koedinger@cmu.edu

Vincent Aleven
Human-Computer Interaction Institute
Carnegie Mellon University
aleven@cs.cmu.edu

ABSTRACT

Although ITSs are supposed to adapt to differences among learners, so far, little attention has been paid to how they might adapt to differences in how students learn from help. When students study with an Intelligent Tutoring System, they may receive multiple types of help, but may not comprehend and make use of this help in the same way. To measure the extent of such individual differences, we propose two new logistic regression models, ProfHelp and ProfHelp-ID. Both models extend the Performance Factors Analysis model (Pavlik, Cen & Koedinger, 2009) with parameters that represent the effect of hints on performance on the same step on which the help was given. Both models adjust for general student proficiency, prior practice on knowledge components, and knowledge component difficulty. Multilevel Bayesian implementations of these models were fit to data on student interactions with a geometry ITS, where students received on-demand problem-relevant help ranging from first-level hints that facilitate application of principles to specific and immediately actionable bottom-out hints. The model comparison showed that in this dataset students differ in their individual hint-processing proficiency and these differences depend on hint levels. These results suggest that we can assess specific learning skills, e.g., making sense of instructional text, and in future work we may be able to remediate and improve such skills.

Keywords

Effect of help on performance, individual differences, learning skills, multilevel Bayesian models, Item Response Theory

1. INTRODUCTION

In virtually all imaginable learning settings, when students work through problems, they may seek help. But are all students able to benefit from help equally, and are there meaningful differences across types of help?[1]

Our long-term goal is to answer this and other questions related to the nature of the learning skills that students bring to bear when working with educational technologies, as well as whether or not there are significant individual differences in these learning skills. Seeking help and learning from help [1, 19] may be one set of such learning skills, which can include both the metacognitive monitoring needed to determine when soliciting help benefits learning, as well as making sense of instructional text in the context of problem solving. If individual differences in learning skills exist, and if they can be assessed, an Intelligent Tutoring System may be able to adapt to these differences, to provide students with appropriate metacognitive support, and perhaps even to improve learning skills.

[1] This work is supported in part by Postdoctoral Training Grant awarded to Carnegie Mellon University by the Department of Education (#R305B110003).

In this project, we aim to determine whether or not there are significant variations in students' abilities to make use of help. As a first step, we examine how well students can use help to solve the task at hand (i.e., the problem step they are working on). While the effect of help on learning, rather than performance, is of primary long-term interest, if a student cannot make good use of help "locally" (on the current step), it is unlikely such help will enhance learning (i.e., enhance performance on a future related task). [11] In other words, studying the "local" effect of help on performance is useful, because any beneficial effect of help on performance may be a harbinger of longer-term effect on learning.

Specifically, our research questions are: How well do students perform after receiving hints, and does performance after hints differ across hint levels? Are there individual differences in how effective hints are among students, and if so, are the individual differences consistent within each student across hint levels? Are the individual differences, if any, related to general student proficiency in solving problems?

We analyze data generated in the course of another study, and use statistical methods to account for potentially confounding variables, including general student proficiency, prior practice on knowledge components, and knowledge component difficulty.

One prior effort to evaluate the effect of hints on same-item performance is by the developers of the Mastering physics ITS. In [12], a 2PL Item Response Theory model was fit to performance on first attempts, after which separate models were fit to each of several paths through the ITS. Unlike that effort, our work examines individual differences with various types of help, and addresses potential confounds due to variability in prior practice and due to difficulty of knowledge components (rather than just unique problem items). We also analyze a larger dataset, and fit parameters relating to various types of help simultaneously in a Bayesian Markov Chain Monte Carlo (MCMC) framework to account for uncertainty during estimation.

Mining data from the Geometry Cognitive Tutor (an earlier version of the tutor whose data is analyzed in the current study), we showed that asking for help is beneficial for local performance. [1] Specifically, asking for help after one or two errors on a step was compared to attempting to solve the step again. Asking for help, compared to continued trying, was associated with fewer subsequent errors on the given step and a reduction in the time needed to complete the step. However, [1] did not look into individual differences in students' ability to take advantage of help to improve performance on problem steps, and did not investigate differences between hint levels.

Another related study [7] presents two models, a learning decomposition and an extension to Bayesian Knowledge Tracing. The latter is particularly interesting in that it aims to distinguish the effect of help as a performance scaffold from its effect on learning. However, neither model addresses multiple hint levels or individual differences in hint-processing proficiency.

Table 1: Examples of hint messages

Knowledge Component	First Hint	Second Hint	Third Hint
Triangle-Sum-Answer	In this problem, you have triangle SOL. You know the measure of two of the angles in this triangle, namely, angles DSO and OLD.	The sum of the measures of the interior angles of a triangle is 180 degrees.	m∠SOL = 180 - m∠DSO - m∠OLD.
Triangle-Sum-Reason	In this problem, you have Triangle WAR. You know the measure of two of the angles in this triangle, namely, angles ARO and OWA.	The sum of the measures of the interior angles of a triangle is 180 degrees.	You can find the measure of Angle WAR by applying the "Triangle Sum" theorem.
Separate-Complementary-Angles-Answer	The problem statement says that angles ∠XSD and ∠JNT are *complementary angles*.	Complementary angles are angles whose measures add up 90 degrees.	m∠XSD = 90 - m∠JNT.
Angle-Addition-Answer	Angles DGF and MGD are *adjacent angles*. This means that they share a side (namely, GD) but do not overlap. Together they form ∠MGF.	When an angle is formed by two or more adjacent angles, the measure of that angle is equal to the sum of the adjacent angles. Therefore, m∠MGF = m∠DGF + m∠MGD.	[No third level hint.]

An exploratory analysis of our dataset (Section 2) shows that selection effects confound a naïve approach that merely tallies successful and unsuccessful performance with and without hints. Section 3 proposes two logistic regression models that take these confounds into account. Section 4 describes the results of fitting multilevel Bayesian implementations of these models to our data. The final sections discuss the results, limitations, contributions of this research, and future directions.

2. EXPLORATORY DATA ANALYSIS

The study that produced the dataset analyzed here took place at a vocational school [17]. Three 9th grade classes of 51 participating students, led by the same teacher, used Geometry Cognitive Tutor as part of regular instruction about twice a week for five weeks. Students worked through problems, most of which contained multiple steps. There were 170 distinct problems, consisting of 1666 problem steps. Problems were assigned to students according to a mastery criterion based on the Knowledge Tracing [8] algorithm in the Cognitive Tutor software, i.e., each student only saw a subset of the 170 problems.

Using this software, a student may make multiple attempts to complete a problem step. Completing a step requires a correct response; giving a correct response on the first attempt means that this student will never see a hint. On each attempt, a student may supply a correct answer, an incorrect answer, or may ask for a hint. The first hint that the student sees is called "help level 1", the second is "help level 2", and so on to the final ("bottom-out") hint, which in our dataset is help level 3 or 4. (Table 1) For students who do not know how to respond, the bottom-out hint often states exactly what the response must be.

In general, a first hint points out relevant problem features, and it defines key terms, e.g., "vertical angles." Second hints state the problem-solving principle that is applicable given the features pointed out in the first hint, in terminology consistent with the first hint. Third hints derive an expression for the sought angle measure (in terms of known angle measures). Using this expression, the angle measure can be found in a straightforward manner, by first substituting in the values for the angle measures referenced in the expression, and then evaluating the resulting arithmetic expression. The rationale for sequencing hint levels from less specific to more specific was to try to give the student as much opportunity as possible to "generate" the step, which may include retrieving a relevant problem-solving principle, as discussed in [4] and [3].

Interaction with such hint sequences may lead some students (e.g., those who are relatively less proficient) to request hints more often than others. Similarly, some problem steps (e.g., those that are challenging) may lead to hint requests relatively more often.

As a measure of student proficiency, we consider how often a student responds correctly to a problem step on the first attempt. Specifically, a crude measure of proficiency is the success rate on first attempts, i.e., the proportion of all problem steps that the student answered correctly on first attempt out of all those first attempts where a student gave a correct or an incorrect response (omitting first attempts where the student requested a hint).

Given this measure, is proficiency related to use of hints? For each student, the hint use rate is the proportion of problem steps on which this student requested one or more hints out of all attempted problem steps. The correlation of student proficiency and hint use rate is r=-0.84, i.e., hints are more likely to be requested by less proficient students.

Similarly, as a measure of problem step difficulty, we take the proportion of first attempts on the step to which a student gives a correct response out of correct and incorrect first attempts (again, omitting first attempts that are hint requests). Is step difficulty related to use of hints? The rate of hint use on a problem step is the proportion of students who request any hints on the step out of all students who attempt the step. The correlation of step easiness (1 - step difficulty) and rate of hint use is r=-0.68, i.e., hints are more likely to be requested on steps that are harder.

Do hints of different levels differ in their effect on performance? If requesting a hint counts as unsuccessful performance (Table 2, top row), the success rate drops from first attempts (78%) to attempts after first and second hints (21% and 37%). However, when students request a first hint, the next action that they are most likely to perform in the tutor is to ask for a second hint (87% of the time). Students ask for a third hint as the likely next action after the second (88% of the time). Not counting hint requests (Table 2, bottom row), performance after the first hint (68%) is lower than after the second and third hints (83% and 88%).

Table 2: Success rates after hints, counting Correct, Incorrect, and Hint outcomes

Success Rate Formula	On First Attempt	After 1st Hint	After 2nd Hint	After 3rd Hint
$C/(C+I+H)$	78%	21%	37%	82%
$C/(C+I)$	83%	68%	83%	88%

To sum up this exploratory analysis, we find that hints are more likely to be requested by less proficient students; hints are more likely to be requested on steps that are difficult; and success after first hints is less likely than after second and third hints.

The exploratory analysis is appealing, but possibly misleading. First, what is "student proficiency"? A student who is proficient may simply have had more opportunities to practice the relevant skills, which would cause a selection effect for this analysis, or there may be additional differences in student ability that cannot be observed directly. Second, while an ITS may tutor all students on the same skills, it may assign students different problems. If so, skills rather than problem steps would be the right grain size for analysis. Third, since students see different problems, and problems involve different hints, there could be selection effects in terms of how we measure performance after hints for different students. Thus, it would be desirable to control for proficiency, prior practice, selection effects related to problem difficulty, and to take into account a model of skills in the domain. As described in the following section, we can use a logistic regression to take these elements into account.

3. METHODS

We fit two models to these data, both extending the Performance Factors Analysis (PFA) model. [14] PFA is a logistic regression that is fit to correct and incorrect student responses.

$$logit(\Pr(Y=1)) = \sum_{j \in KC} (\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j})$$

Equation 1: Performance Factors Analysis (PFA) model

Under PFA, the probability of a correct response by a pupil on a problem step, i.e., of $Y = 1$, is determined by a linear combination of parameters related to the knowledge components (KCs) that are thought to be relevant to that step. Parameter β_j denotes the easiness of KC_j. Parameters γ_j and ρ_j are weights on the observed frequency of successful ($s_{i,j}$) and unsuccessful ($f_{i,j}$) prior practice by the same learner i on the same KC j. The innovation in PFA was to separate γ_j and ρ_j, the effects of successful and unsuccessful prior practice, rather than collapsing these effects as one parameter.

Table 3: Example of instances in our dataset

Pupil	Item	Attempt	Prior Practice	Outcome
5	Prob1.St3	1	$S_{5,9}$=3; $F_{5,9}$=1	First hint
5	Prob1.St3	2	$S_{5,9}$=3; $F_{5,9}$=1	Incorrect
5	Prob1.St3	3	$S_{5,9}$=3; $F_{5,9}$=1	Correct

Our interest is in learner performance in the presence of help on attempts after the first. The original use of PFA was to model unassisted performance; in PFA, the outcome variable Y and the prior practice counts $s_{i,j}$ and $f_{i,j}$ only represent first attempts on a problem step, not subsequent attempts. By contrast, we fit our models to outcomes both at first attempts and at each attempt that was the next action after a hint (but the prior practice counts still represent only first attempts).

Consider the example in Table 3, where a student (pupil 5) makes three attempts on the same item (problem 1, step 3). When the ITS initially presents the student with this step in the course of solving the problem, the student requests a hint. This hint is at the first of several levels of help (usually 3 or 4) that the ITS may offer on a problem step. According to the knowledge component model for the problems in this dataset, this step has a single relevant KC with KC id=9. This student has had prior practice opportunities with this KC: three were successful, and one was not. Counts of prior practice are based on first-attempts only; thus, when this student practices this KC on a future item, prior practice counts will be $S_{5,9}$=3; $F_{5,9}$=2, because the outcome of the first attempt in this example was unsuccessful. This example yields two instances to be input to the logistic regression, corresponding to the first 2 attempts. Both attempts are coded as having the outcome 0 (only correct outcomes are coded as 1). For the purpose of estimating the help-level parameters in our model, the first-attempt instance is coded as not following a help message, and the second-attempt instance is coded as following a hint at help level 1. We assume that the effect of a hint should be observable in the next attempt on the same step. Because attempt 3 follows an input rather than a hint display, its outcome is not directly attributable to a hint, and this attempt does not yield an instance. Of the 28777 transactions in this dataset, 17515 were first attempts, 4466 attempts were the next action after some kind of a hint, and the rest were not entered as instances because they were not next actions following a hint.

The first model, ProfHelp, examines how help levels differ in their effect on performance, but does not consider individual differences in hint processing among students.

$$logit(\Pr(Y=1)) = \theta_p + \lambda_h + \sum_{j \in KC} (\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j})$$

Equation 2: Proficiency and Help (ProfHelp) model

The innovation in this model is the λ_h parameter. One λ_h is fit for every attempt after a hint. (Because help may be requested as a first attempt, but never prior to a first attempt, $\lambda_0 = 0$.) One of $\lambda_1, .., \lambda_4$, respectively, represents the contribution of having just seen a first, second, third or fourth hint to the probability of successful performance on this subsequent attempt. Another view of λ_h is that it represents average proficiency in processing level-*h* hints. Parameters other than λ_h control for student proficiency, problem step difficulty via a decomposition on knowledge components, and prior practice on knowledge components. In other words, the effect of having just seen a hint is not confounded by the findings that hints are more likely to be requested by less proficient students and on more difficult items (Section 2), nor by the intuition that a lack of prior practice can lead to more frequent hint requests. Finally, one θ_p parameter, as in Item Response Theory (IRT) models, is fit for every student *p*, representing the baseline proficiency of that student.

$$logit(\Pr(Y=1)) = \theta_p + \lambda_{p,h} + \sum_{j \in KC} (\beta_j + \gamma_j s_{i,j} + \rho_j f_{i,j})$$

Equation 3: ProfHelp-ID (Individual Differences) model

The second model, ProfHelp-ID, considers that the same help level may have different effects on different students. The

difference from the ProfHelp model is in the $\lambda_{p,h}$ parameter, where the subscripts p, h indicate that a separate parameter is fit for each pupil in each help level. This represents the pupil's individual hint-processing proficiency. These parameter estimates are pooled across pupils within a single help level via a multilevel model (bold typeface denotes hyperparameters):

$$\lambda_{p,h} \sim N(\boldsymbol{\lambda}_h, \boldsymbol{\sigma}_h^2)$$

For instance, the ProfHelp-ID model stipulates that $\lambda_{p,2}$, i.e., each per-pupil estimate of the effect of responding after a second hint (h=2) is drawn from a distribution with mean $\boldsymbol{\lambda}_2$ and variance $\boldsymbol{\sigma}_2^2$ that is shared across pupils. In this way, information on each pupil helps determine a baseline effect of seeing a second hint, and the baseline effect helps constrain the estimate of the per-pupil individual differences.

Partial pooling is appropriate for this problem not only for statistical parsimony, but also because it lets us be conservative in making a claim about the presence of individual differences. (Partial pooling is similar to the idea of a random effect, where values are assumed to come from a broader sample of interest, rather than a fixed effect, where all values of interest are represented.) The alternative, a no-pooling model, would treat pupils as independent of one another. This means that first, the no-pooling model could detect individual differences even when the differences are small (i.e., not meaningful), and second, unpooled individual differences would be hard to quantify because there may be very few observations for any particular pupil at a given help level. The partial pooling pulls all individual difference estimates towards the mean, reducing the effect of small differences, and it helps compensate for data sparsity by using the hyperparameter estimates as prior information for the parameters. (Note that model ProfHelp is the complete-pooling version of ProfHelp-ID, in that ProfHelp does not allow for individual differences in hint processing.)

The models were fit using the JAGS software for Bayesian modeling [15], which is an effective platform for fitting Item Response Theory and similar models (e.g., [9]). For each model, we ran 4 sampling chains, with 400 adaptation iterations (discarded). Inferences below are based on every 10th draw (thinning) of 1000 iterations. Model convergence and mixing across chains were verified by visual examination of autocorrelation, trace and density plots.

4. RESULTS

As multilevel Bayesian models, ProfHelp and ProfHelp-ID may be compared in terms of Deviance Information Criterion (DIC). DIC is similar to AIC in that it rewards models that fit the data well but penalizes an increase in the number of parameters in the model. [16] DIC takes into account that in Bayesian models with pooling, the effective number of parameters is itself estimated as a posterior distribution of a random variable.

Table 4: Model-fitting results

Model	Deviance	Effective Parameters	DIC
ProfHelp	22013	135	22149
ProfHelp-ID	21741	220	21962

As Table 4 shows, the ProfHelp-ID model is preferable to the ProfHelp model on this dataset in that the improvement in prediction accuracy outweighs the increase in the number of parameters. Relative to the ProfHelp approach of fitting a single parameter across all students within a help level (complete pooling), the partial-pooling approach of ProfHelp-ID finds that there are individual differences in student performance after a hint at each help level. This finding is despite the fact that ProfHelp-ID is nonetheless more conservative than a no-pooling model.

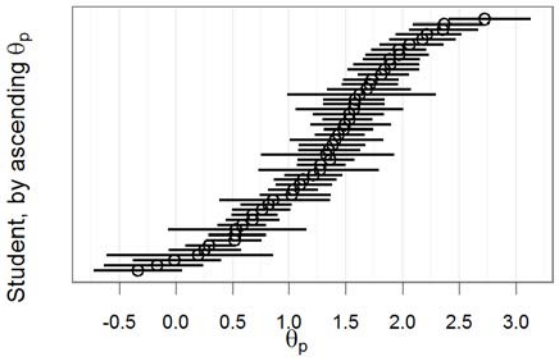

Figure 1: Medians and 95% CI for θ_p under ProfHelp-ID

The θ_p proficiency parameter (Figure 1) is positive for most of the students, reflecting the prevalence of successful first attempts in this dataset (the model predicts that a student for whom $\theta_p = 0$ will answer correctly on 50% of first attempts, given that the other terms are zero). The θ_p parameter is entered into the model for both first attempts and later attempts, and both could affect its estimate. However, first attempts are much more frequent than later ones, and $\lambda_{p,h}$ provides an intercept for each pupil on the later attempts. This effectively makes θ_p a constant baseline for $\lambda_{p,h}$ that is unaffected by the later attempts.[2]

ProfHelp-ID measures the effect of having seen a hint on the immediately preceding attempt as a baseline across all students (the λ_h hyperparameter), and as a deviation from this baseline for every pupil, $\lambda_{p,h}$. The improved fit of ProfHelp-ID over ProfHelp implies that the mean effects λ_h are correct only on average, not for all students. As Figure 2 shows, the mean effect, in logit units, of having just seen a hint (solid black vertical line in each of the three frames) are approximately -2.4, -1.7 and 0.5 for first, second and third help levels, respectively.[3] These differences are significant, as indicated by the non-overlapping 95% credible intervals (grey vertical lines on the left and right of each black line). The mean effects of first and second hints are negative, which implies that, on average, the performance of all students, proficient or not, and on all problem steps, easy or difficult, is lower after these hints than would be predicted based only on overall proficiency θ_p. The effect of third hints is only somewhat

[2] A parameter in a logistic regression adds to the model's estimate of success on a given instance. To interpret a coefficient, a rule of thumb is to divide by 4. For example, if $\theta_1 = 2$, that adds 0.5 to the probability that model will predict success on every attempt by pupil 1.

[3] As a check on the model fitting, the estimates of $\lambda_{p,h}$ from ProfHelp were similar, -2.3, -1.5, and 0.5. There were few observations for performance after a fourth hints, so we omit discussion of $\lambda_{p,4}$ and λ_4.

positive. Converted to probabilities, effects of first and second hints at -2.4 and -1.7 logit units, respectively, implies that a student with median proficiency on this dataset ($\theta_p = 1.4$), on a problem step of average difficulty ($\sum_{j \in KC} \beta_j = 0$), and with no prior practice on relevant KCs, is predicted to respond correctly 27% of the time after first hints and 42% of the time after second hints. These predicted correctness rates are higher than those of the "naïve" analysis (21% and 37%, Table 2) that does not take into account proficiency and other confounds. While these rates are low, they are nonetheless an improvement over the students' failures to answer correctly on the first attempt.

An unexpected finding is that general proficiency θ_p is *negatively* correlated with hint-processing proficiency $\lambda_{p,h}$: for first, second, and third hints, r=-0.48, r=-0.54, and r=-0.41, $p < 0.01$ for all. The more proficient the student, the less likely it is that the student benefits from a hint. This relationship is also visible in Figure 2, where each frame is ordered by ascending θ_p. Hint-processing proficiency of first hints $\lambda_{p,1}$ is also correlated with hint-processing proficiency of second hints $\lambda_{p,2}$, r=0.34, p<0.05; other hint-processing proficiencies are uncorrelated with each other.

5. DISCUSSION

We aimed to understand the nature of learning skills such that we can support learning more effectively. We found that hints levels differed in their effect on performance, and only level-1 and level-2 hint-processing proficiencies correlated with each other. Further, there were individual differences in hint-processing proficiency, and general proficiency was negatively correlated with hint-processing proficiency.

Given how hint levels are implemented (Table 1), it is not surprising to see better performance on the next attempt after the bottom-out hint level, compared to the next attempt after other hint levels. As mentioned, all that correct performance following a bottom-out hint requires is algebraic substitution and arithmetic, which are likely to be mastered skills for our student population. By contrast, correct performance after first and second hint levels requires interpretation of mathematical text that refers to potentially unmastered geometry concepts and principles. To solve problems in the geometry unit in this dataset, one needs to retrieve a general geometry principle, to apply the principle to the problem by mapping it to specific problem features, and to perform algebra and arithmetic according to the principle. Before the principle can be retrieved, salient problem features need to be identified. Level-1 hints tend to point out the salient problem features and define key terms. Level-2 hints state what principle is applicable given the salient features pointed.

The negative effects of level-1 and level-2 hints are consistent with prior work on hint effectiveness. [1] As pointed out in [7], "students request help on [items] on which they have low knowledge. The help thus acts as *evidence* of a lack of knowledge, rather than a direct *cause* of that lack of knowledge." Further, neither short nor long hint reading times are positively associated with learning. [18] Another explanation for the negative coefficients for our dataset in particular is that the logistic regression is effectively forced to estimate these very negative effects given the prevalence of positive θ_p values (which are in turn due to the prevalence of successful first attempts).

Prior work suggests that it can be fruitful to consider how tutor behavior may differentially affect students across varying levels of KC mastery. [2] The ProfHelp models are based on the

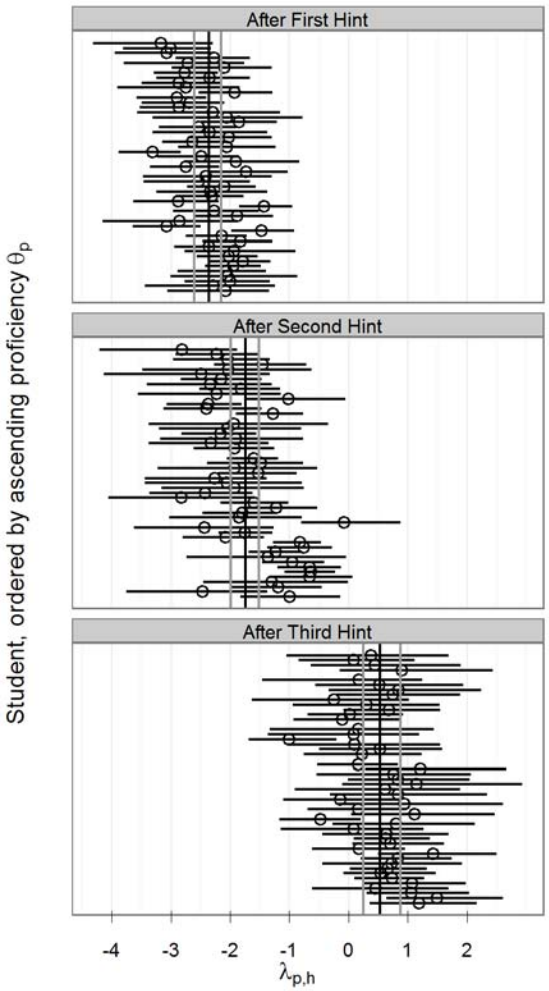

Figure 2: Medians and 95% CI for $\lambda_{p,h}$ under ProfHelp-ID; median of λ_h (black vertical) and 95% CI (grey vertical)

psychometric concept of latent traits that is inherent to Item Response Theory. IRT models are said to be unidimensional if they represent proficiency with one parameter θ_p per student. ProfHelp-ID relaxes this unidimensionality assumption via parameters $\lambda_{p,h}$ per student for attempts after hints, but retains it within each type of attempt (first attempt and after each hint level). Thus, the dimensions of proficiency in ProfHelp-ID (first attempts and help levels) may not represent proficiency ideally. The ProfHelp-ID estimate of the probability of success will be in error when performance within this type of attempt is multidimensional, e.g., if an otherwise easy KC unexpectedly challenges a proficient student (or, vice versa, if a student with low proficiency succeeds quickly on a generally difficult KC). Having found individual differences within different attempt types, we speculate as to the nature of the learning skills that may be involved in interpreting hints and using them to support correct performance. This analysis will inform future model refinements.

Success after level-1 hint with good knowledge of relevant KC. A student who is close to KC mastery did not succeed on the first attempt on a step, but did on the next action after a level-1 hint. The failure on the first attempt may have been an "identification

slip", i.e., a slip in identifying the relevant problem features that was due to random circumstance rather than lack of knowledge, or to high cognitive load such as could be expected in a dataset of quite complex geometry problems that involve multiple steps and multiple problem-solving principles. Level-1 hints point out problem features that are relevant to the application of a principle, but not what principle to use, or how. Because the student succeeded after the level-1 hint, the student was apparently able to retrieve and apply the principle (i.e., did not need further hints) once given the salient features, but required assistance to identify the salient features. When hints are used to fix "identification slips," no hint interpretation skills are needed; the hint serves as reminder of something the student already knows but failed to retrieve. The student still applied "principle application skills" to the extent that the knowledge of how to apply this principle had not yet been proceduralized or automated.

Success after level-1 hint with little knowledge of relevant KC. By contrast, an identification slip is not possible for a student with little knowledge of the relevant KC. Given that level-1 hints state problem features relevant to the application of a principle, success after a level-1 hint suggests that this hypothetical student was able to infer a correct answer from a set of problem features, even without knowing the rule that connects the features to the answer. Perhaps Assuming this was not a guess, the student induced a valid principle from the given example, and then used principle application skills mentioned above, though a less generous interpretation would suggest that the student learned shallowly. Quite an impressive feat of unsupervised inductive learning, with less than a single example to work with and no outcome given! How could this be possible? Perhaps this student drew on additional information sources, e.g., student peers or the textbook. Perhaps the diagram helps; e.g., once one sees a visual representation of adjacent angles, the notion that the measure of an angle made up of adjacent angles is the sum of the two measures of the adjacent angles seems quite intuitive. Other geometry knowledge may help as well. For instance, smart students may be able to infer the vertical angles theorem from the linear pair postulate.

Success after level-2 hint with good knowledge of relevant KC. Failure after a level-1 hint followed by success after a level-2 hint suggests that the student needed to be reminded of the relevant domain principle. This student should have been able to retrieve the relevant domain principle from memory given the prior practice of the KC. What could cause failure to retrieve a principle? Similar to failure on a first attempt, one cause may be a mere "applicability slip" in mapping problem features to a known principle, e.g., due to random occurrence or to overwhelming cognitive load. Another cause may be that there is a phase in the normal skill acquisition process in which students have more trouble recognizing the applicability of rules than in applying them once cued to critical problem features. In other words, while in this phase, students need to be reminded of a principle, but can apply it, especially when also given some key information (as in the level-1 hint) on how to instantiate the principle. This hypothesized phase also explains failure after the level-1 hint.

The modest but statistically significant correlation of $\lambda_{p,1}$ and $\lambda_{p,2}$ suggests that the two hint levels may be linked in how they affect students, but that there are also some differences. One explanation for the correlation is that level-1 and level-2 hints would both be skipped by a student engaged in "help abuse" [19], causing both level-1 and level-2 hints to be associated a 0 (unsuccessful) logistic regression outcome. By contrast, bottom-out hints cannot be skipped, so unsuccessful outcomes after bottom-out hints would not be confounded with help abuse. Another cause for the correlation may well the requirement, shared across the level-1 and level-2 hints, to apply a principle, while the requirements of bottom-out hints, likely mastered by all students, would not induce a correlation. Finally, the two hint levels may share the hypothesized phase affecting students with good KC knowledge.

One difference between level-1 and level-2 is that answering correctly after (only) a level-1 hint requires more domain-specific knowledge than answering correctly after a level-2 hint. One way to answer correctly after a level-1 hint is to retrieve the relevant problem-solving principle from memory, possibly cued by the problem features pointed out in the hint, and to apply the rule successfully, helped perhaps by the information provided in the hint. By contrast, to answer correctly after a level-2 hint, it is not necessary to retrieve the principle from memory, since the level-2 hint provides a statement of the principle. The student must still do some work to figure out how the rule applies.

An instance of poor retrieval may be symptomatic of a broader retrieval deficiency on the part of the student, which would constitute a learning skill deficiency. Success after first hints occurred frequently enough (predicted 27% correctness rate for a student with median general proficiency) that it may be worth investigating whether such a deficiency could be detected, or even addressed. Ideally, learners could be supported in overcoming such a cognitive shortcoming on their own. Students need to apply general principles to specific problems in many domains (e.g., [10]), and it would be interesting to see if such a skill could transfer.

Success after level-2 hint with little knowledge of relevant KC. Poor retrieval cannot explain success after a level-2 hint when a student has had little prior practice on the relevant KC, i.e., when there is no expectation for retrieval. A level-2 hint states the rule that applies, but not *how* it applies. Thus, such successful performance may indicate that the student is skilled at applying a somewhat unfamiliar problem-solving principle, when given a statement of that principle (level-2 hint) and key problem features that instantiate the principle's applicability conditions (level-1 hint). What remains for the student to do is still rather involved: apart from understanding the principle, the principle has to be mapped onto the problem, a process that (facilitated by the level-1 hint) requires dealing with difficult terminology in relating the general terms in which the rule is stated to the specific problem.

Success after level-3 hint. With respect to the level-3 hints, it seems unlikely that correct performance after a bottom-out hint involves important learning skills, aside from possibly a general tendency to carefully follow very specific instructions.

In sum, this analysis contemplates several hypothesized metacognitive skills. Success after level-1 and level-2 hints for a student with high KC knowledge may indicate deficiencies in identifying salient problem features, mapping a principle to salient features, and retrieving a principle. Success after level-2 hints for a student with little KC knowledge may indicate skill in applying unknown principles (i.e., parsing and mapping—with some help—of an unfamiliar principle). Our results could be viewed as implying that different students possess these different learning skills to different degrees. This interpretation addresses both possible causes of differences between hint levels and possible causes of learners' differences in hint processing. For instance, if

we could find a way to help students learn to recognize when a geometry principle applies, this should both improve the effectiveness of first hints, and reduce unexplained variability among students in terms of their hint-processing proficiency. Perhaps if students were given instruction to look for diagram features that can cue a principle, then on a first hint like "The problem statement says that angles ∠XSD and ∠JNT are *complementary angles*", they might be better able to interpret the notion of complementary angles by paying attention to that part of the diagram. The analysis considered the major findings that hint levels differ in their effect on performance, that student proficiencies with level-1 and level-2 hints are modestly correlated, and that there are individual differences in hint-processing proficiency.

Finally, we address the finding that general proficiency is negatively correlated with hint-processing proficiency. One explanation is that this finding is merely an artifact induced by the statistical model. In designing the ProfHelp-ID model, we reasoned that to ascribe an effect to some proficiency with hints we had to partial out the effect of general proficiency. In fact, making θ_p a baseline for $\lambda_{p,h}$ may overcorrect for any relationship between general proficiency and hint proficiency. The linear combination of the two parameters effectively subtracts θ_p from $\lambda_{p,h}$, which means that $\lambda_{p,h}$ contains information on θ_p, and that can induce the negative correlation.[4] While such a correlation complicates interpretation of parameter estimates, it would not invalidate the model fit. A second explanation is that individuals with a higher proficiency may be less proficient with hints because they have less practice using them.

One contribution of the ProfHelp models is that they control for selection effects due to general student proficiency, prior practice on knowledge components, and knowledge component difficulty. The models here do not account for other selection effects, which we intend to address in future work. First, ProfHelp treats all hint messages at a given level as equally effective, while messages associated with different KCs may in fact have differential effects on student performance. (Such an analysis would be the "KC differences" analogue of the individual differences analysis presented here.) In this way, we might be able to identify specific hint messages that are significantly more or less effective than other messages to inform ITS design. Second, in future work we intend to relax the unidimensional IRT assumption, i.e., to handle the case that a KC that the model estimates to be easy may challenge a student that the model estimates to be proficient. Third, the ProfHelp models do not account for patterns of use that students may follow. For instance, in discussing the effect of level-1 and level-2 hints it would be desirable to account for the effects of help abuse, e.g., a student clicking through the hints without reading them. [2, 6] The ProfHelp models do not distinguish such behavior from spending a long time on each hint, which may indicate deliberative reflection.

The need for future research is highlighted by the ProfHelp-ID estimates of effectiveness of hints: 27% and 42% accuracy after level-1 and level-2 hints, respectively, for a student with median general proficiency. While even these relatively low levels of effectiveness improve, by definition, over the students' failure to answer correctly on the first attempt, there is clearly room to make hints more effective, and hence a need for research on hints types and hint processing. The ProfHelp-ID model may serve as a tool for such research. Given that this model can fit transaction data from an ITS, one can expect to apply it again in the future to evaluate alternative hinting strategies.

6. CONCLUSIONS

The results presented here may be said to pose more questions than they answer, which is appropriate for an early project in a relatively unexplored area. Significantly, the results show that hints levels differ in their effect on performance, and that there are individual differences in hint-processing proficiency. These findings account for general student proficiency, prior practice on knowledge components, and knowledge component difficulty via the ProfHelp and ProfHelp-ID models. The next steps are to understand the causes of the individual differences, and to try to detect them automatically.

An additional contribution of this research is the new Bayesian implementation of the new ProfHelp and ProfHelp-ID models (and by extension, the PFA model).[5] The flexibility of the JAGS modeling tool is well-suited to logistic regressions such as these and to the need for rapid prototyping of model variations. The time saved in development easily outweighs potentially slow MCMC sampling. Moreover, the model-fitting process can easily be parallelized for separate MCMC chains.

This research has wide impact. The data analyzed here come from a system in the Cognitive Tutor family, in use by over 600,000 students. [5] The same methods would apply to any software that uses either progressive hint sequences or multiple independent types of help. For instance, in SQL Tutor, "an error flag message informs the student about the clause in which the error occurred. A hint-type message gives more information about the cause of error. Partial solution feedback displays the correct content of the clause in question, while the complete solution simply displays the correct solution of the current problem." [13] The Masteringphysics ITS includes three types of hints ("a list of steps, declarative statements, and procedural subtasks") and other types of help. [12]

Among the limitations of this research, the first is that it considers the effect of hints on performance, not learning. As [7] points out, in theory, a hint may both scaffold performance on the current step and it may teach the student in preparation for a subsequent problem. However, as evidenced by the analysis in Section 5, while the effects on learning are important, the effects on performance are not yet well understood.

Other limitations are due to the assumptions embedded in the PFA model and the ProfHelp models. These include that knowledge components are independent and linearly additive; that the effects of the problem step are fully represented by the relevant knowledge components and prior practice on these KCs; and that there are no problem effects, e.g., steps within the same problem are treated as independent of one another. The ProfHelp models are limited in that they only consider the effect of help from the immediately preceding attempt, while there could be effects that carry over from earlier attempts. In the dataset examined here, hint levels were always presented in the same order, and the differential effects of hint types could not be teased apart using

[4] To see how this would work, let X and Y be two independent random normal variables. Let $X' \leftarrow Y - X$. By definition, $cor(X,Y) = 0$, but $cor(X',Y) \sim 0.71$.

[5] Please contact the corresponding author for the JAGS code.

ProfHelp. However, this is a limitation of the dataset rather than ProfHelp itself.

In future work, we plan to extend the ProfHelp models. We may incorporate students' hint-level preferences, such as to take into account the tendency of some students to click through to the bottom-out hint without making attempts after first and second hints. We may also incorporate the number of prior hint episodes on practice opportunities of various KCs to distinguish the effect of prior hints from the effect of incorrect prior performance.

At the same time, regression techniques cannot eliminate all selection effects. Future work should include controlled experiments that compare different hint types, and an evaluation of their effects on learning and on reduction of unexplained variance in hint processing among students.

7. ACKNOWLEDGMENTS

The authors thank Brian Junker and Georg Goerg for advice on statistical modeling, and S. McKay Curtis for addressing many questions about the mcmcplots package.

8. REFERENCES

[1] Aleven, V. and Koedinger, K.R. 2000. Limitations of Student Control: Do Students Know when They Need Help? *Intelligent Tutoring Systems* (Berlin, Heidelberg, 2000), 292–303.

[2] Aleven, V., Mclaren, B., Roll, I. and Koedinger, K. 2006. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*. 16, 2 (2006), 101–128.

[3] Aleven, V., Stahl, E., Schworm, S., Fischer, F. and Wallace, R. 2003. Help Seeking and Help Design in Interactive Learning Environments. *Review of Educational Research*. 73, 3 (Sep. 2003), 277–320.

[4] Anderson, J.R. 1993. *Rules of the Mind*. Psychology Press.

[5] Apollo Group to Acquire Carnegie Learning: 2011. *http://carnegielearning.com/press-room/press-releases/2011-08-02-apollo-group-to-acquire-carnegie-learning/*. Accessed: 2012-04-16.

[6] Baker, R.S., Corbett, A.T., Koedinger, K.R. and Wagner, A.Z. 2004. Off-task behavior in the cognitive tutor classroom: when students game the system. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004), 383–390.

[7] Beck, J.E., Chang, K., Mostow, J. and Corbett, A. 2008. Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. *Intelligent Tutoring Systems*. B.P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, eds. Springer Berlin Heidelberg. 383–394.

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

[9] Curtis, S.M. 2010. BUGS Code for Item Response Theory. *Journal of Statistical Software*. 36, 1 (2010), 1–34.

[10] Goldin, I.M., Pinkus, R.L. and Ashley, K.D. accepted. Validity and Reliability of an Instrument for Assessing Case Analyses in Bioengineering Ethics Education. *Science and Engineering Ethics*.

[11] Kirschner, P.A., Sweller, J. and Clark, R.E. 2006. Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educational Psychologist*. 41, 2 (Jun. 2006), 75–86.

[12] Lee, Y.-J., Palazzo, D., Warnakulasooriya, R. and Pritchard, D. 2008. Measuring student learning with item response theory. *Physical Review Special Topics - Physics Education Research*. 4, 1 (2008).

[13] Mitrović, A. 1998. Experiences in implementing constraint-based modeling in SQL-Tutor. *Intelligent Tutoring Systems* (1998), 414–423.

[14] Pavlik Jr, P., Cen, H. and Koedinger, K. 2009. Performance Factors Analysis - A New Alternative to Knowledge Tracing. *14th International Conference on Artificial Intelligence in Education* (Brighton, England, 2009), 531–538.

[15] Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vienna, Austria, 2003).

[16] Plummer, M. 2008. Penalized loss functions for Bayesian model comparison. *Biostatistics*. 9, 3 (Jul. 2008), 523–539.

[17] Salden, R.J.C.M., Aleven, V.A.W.M.M., Renkl, A. and Schwonke, R. 2008. Worked Examples and Tutored Problem Solving: Redundant or Synergistic Forms of Support? *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (Austin, TX, 2008).

[18] Shih, B., Koedinger, K. and Scheines, R. 2010. Unsupervised discovery of student learning tactics. *Proceedings of the Third International Conference on Educational Data Mining* (2010).

[19] Wood, H. and Wood, D. 1999. Help seeking, learning and contingent tutoring. *Computers and Education*. 33, 2 (1999), 153–170.