Hints: You Can't Have Just One

Ilya M. Goldin Kenneth R. Koedinger Vincent Aleven
Human-Computer Interaction Institute
Carnegie Mellon University
goldin@cmu.edu koedinger@cmu.edu aleven@cs.cmu.edu

ABSTRACT

A student using an interactive learning environment (ILE) may take multiple attempts to solve a problem step, at times using hints. But how effective are hints? Because data mining occasionally finds implausible negative effects of hints, a method is needed to remove selection effects related to hint use.

We distinguish multiple attempts in which a student repeatedly seeks hints from multiple attempts to answer the problem. Exploratory analysis of log data from a tutoring system shows that making a hint request rather on the first attempt on a problem step correlates with hint requests on subsequent attempts, and proficiency on a first attempt correlates with proficiency on subsequent attempts. Based on this, we devise a multinomial logistic regression that distinguishes hint-request tendency from proficiency. We find that seeking just one hint is associated with repeated hint-seeking, but when students do make attempts to solve a problem after viewing a hint, they succeed about half of the time. Thus, the model removes seemingly negative "effects" of hints. We also find that individual differences among students are more prominent in hint-seeking tendency than in proficiency with hints. We conclude with some ideas to improve our model.

Keywords

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

1. INTRODUCTION

Work on help-seeking in Interactive Learning Environments (ILEs) shows that effects of help are not always straightforward. For instance, different types of hints may differ in effectiveness, and students may differ in proficiency with hints [5, 6]. Further, students may have a variety of help-seeking behaviors, such as help-avoidance (a failure to seek help when the student would likely benefit from it), and help-abuse (seeking help when the student can likely answer the problem). [1] Occasionally, use of help may be linked with *negative* effects [2, 4, 5], but the negative estimate is unsatisfactory. It is doubtful that hints *cause* incorrect performance: although a hint may at times confuse a student and thus contribute to an error, it does not reduce student knowledge. More plausibly, a hint request evidences that the student has not understand the material. In a sense, negative estimates of hint effects imply that the statistical method behind these estimates is a poor representation of human performance (or learning). A better model would reflect a positive or neutral hint effect.

We consider whether the negative hint effects estimated in prior work are due to student tendency to request multiple hints without intervening attempts that could solve the problem. For example, one perspective is that attempt outcomes are effectively binary indicators of skill mastery, either successful (if correct) or not (if incorrect or a hint request). An incorrect attempt suggests that skill mastery is somehow deficient (although it may also be a slip), and a hint request suggests that the student does not know enough to answer the problem. Nonetheless, students may request hints to learn. Because hint requests and incorrect attempts can differ, we may need to distinguish tendency to answer correctly or incorrectly (proficiency) from tendency to request hints.

While we can only hypothesize about the myriad reasons that students may have to avoid or overuse hints, something we can quantify is a tendency to request a hint rather than answer incorrectly, i.e., when a hint may help [8]. This Tendency to Ask for Help-Not Risking an Incorrect (TAH-NRI) may differ across categories of attempts; for instance, on average, students may like to try to solve a problem a second time rather than to use hints. Further, students may differ in their tendency to request hints.

Counts of student actions may underestimate TAH-NRI, e.g., if the student answers incorrectly because a hint was unavailable, and overestimate it, e.g., if the student only seeks the bottom-out hint and must skip other hints to get to the last hint in a sequence. Proficiency may be viewed as the correctness rate when the student actually gives an answer. An operational definition of that is declining to request a hint, but students decline for a variety of reasons, including when they actually desire help. For instance, a student may be aware of own lack of prerequisite knowledge, yet may have poor experience with hints. In this light, distinguishing TAH-NRI from proficiency is a crude but potentially useful representation of metacognition.

We examine these two notions empirically. First, we explore frequencies and correlations of student proficiency and TAH-NRI on different types of problem-solving attempts. Second, this exploratory analysis informs a statistical model of proficiency and TAH-NRI. The model improves on exploratory estimates of proficiency and TAH-NRI by taking other predictors into account.

2. EXPLORATORY DATA ANALYSIS

We perform exploratory and model-based (Sec 0) analyses on a dataset of 51 9th grade students using the Geometry Cognitive Tutor. The students worked through 170 geometry problems, consisting of 1666 problem steps (about twice a week for five weeks). Each student only saw a subset of the 170 problems. In the Geometry Cognitive Tutor, a student may make multiple attempts to complete a problem step. Completing a step requires a correct response. On each attempt, a student may supply a correct answer, an incorrect answer, or may ask for a hint. We omit second hint displays; a hint's effectiveness in a specific problem step for a specific student is only evaluated once.

In our analysis, we first consider that how students behave on a first attempt on a problem step yields contextual information for understanding subsequent attempts. This leads us to explore the relation of first-attempt hint-request behavior to behavior on subsequent attempts. Second, we ask whether there are individual differences among students in hint requests and proficiency that may characterize the behavior of a student across time.

We group hint messages that differ in terms of surface features, i.e., in terms of the names and measures of the angles in a geometry problem. [6] We manually categorize each group of hint messages as feature-pointing, principle-stating or providing the

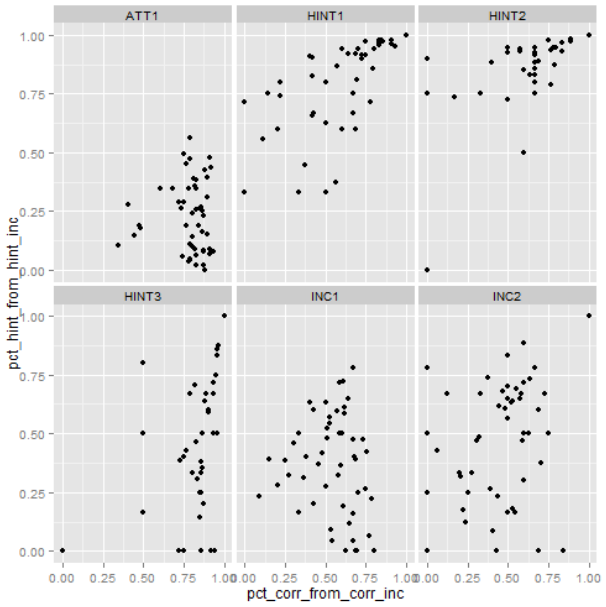

Figure 1: Average student proficiency vs. average TAH-NRI on each attempt type (in percentage-based measures). From top left: first attempts, attempts after feature-pointing hints, after principle-stating hints, after bottom-out hints, after the first incorrect outcome, after the second incorrect outcome.

bottom-out hint. Feature-pointing hints make salient important problem features, e.g., by pointing out that two particular angles are vertical angles. Principle-stating hints give a domain-specific principle that is necessary to solve the problem, e.g., that vertical angles are equal in measure. Bottom-out hints show how to find the answer to the problem, such as by summing known quantities.

Table 1: Rates of hint request and incorrect outcomes

First-Attempt Rate			
Hint Requests		Incorrects	
5%		16%	
Second-Attempt Rate		**Second-Attempt Rate**	
Hint Requests	Incorrects	Hint Requests	Incorrects
70%	5%	22%	35%

How does first-attempt hint behavior relate to behavior on a subsequent attempt? On a first attempt to solve a problem, students request hints only 5% of the time (Table 1), and enter incorrect responses more often (16%). On a second attempt, the conditional probability of a hint request given that the first attempt was also a hint request increases from 5% to 70%. In other words, students are unlikely to request a hint in the first place, but there is an extremely high rate of second hint requests after a first hint, 70% of all second attempts and 93% [70/(70+5)] of the non-correct second attempts after a hint. Statistical models of help-seeking behavior should take first-attempt behavior into account.

Second, we consider individual differences. In exploratory analysis, we define proficiency as the percent correct out of all attempts where a student *p* actually tries to solve the problem, i.e., attempts may be correct or incorrect, but not hint requests: $\frac{Corr_p}{Corr_p+Inc_p}$. Similarly, we define a student's TAH-NRI as the percent of hint requests out of all attempts when the student likely does not know enough to solve a problem, i.e., attempts that end in a hint request or incorrect, not in a correct outcome: $\frac{Hint_p}{Hint_p+Inc_p}$.

Figure 1 presents a distribution of proficiencies (X axis) and hint-request tendencies (Y axis) on each attempt type. Individual differences in proficiency and in TAH-NRI characterize a student across opportunities on each attempt type. Proficiency on attempts after hints is moderately or strongly related to TAH-NRI on those types of attempts. Proficiency on attempts after a first incorrect is weakly or even negatively related to TAH-NRI, but proficiency after a second incorrect is moderately related to TAH-NRI.

A student who is proficient on one attempt type should also be proficient on others. We find (Table 2) that proficiency on first attempts is moderately ($r = 0.39$) related to proficiency on attempts after a feature-pointing hint and after a second incorrect outcome ($r = 0.49$), and strongly related to proficiency on attempts after the first incorrect ($r = 0.74$). Nonetheless, proficiency on first attempts is not related to proficiency after other hints, nor to TAH-NRI. In general, we should take first-attempt proficiency into account when predicting performance on other attempts, but not necessarily when predicting hint requests.

Table 2: Correlations of first-attempt proficiency with proficiency and TAH-NRI on other types of attempts

Attempt Type	Corr. vs. Proficiency on Other Attempts	Corr. vs. TAH-NRI
First attempts	1.00	0.01
After FP Hint	0.39	0.12
After PS Hint	0.18	-0.08
After BOH	0.20	0.47
After 1st Incorrect	0.74	0.14
After 2nd Incorrect	0.46	0.30

In sum, models should account for student differences, and student effects are different when predicting a hint request rather than a correct outcome.

3. MODELING

We present a baseline-category multinomial logistic regression to predict which outcome is most likely on an attempt (*INCORRECT*, *HINT-REQUEST*, *CORRECT*). *INCORRECT* is the baseline outcome against which other outcomes are compared. With K outcomes, there are $K-1$ comparisons: comparison $k = 1$ of *INCORRECT* vs. *HINT-REQUEST* yields parameters related to TAH-NRI; comparison $k = 2$ of *INCORRECT* vs. *CORRECT* yields parameters related to proficiency.

As a basis for this multinomial model, we take the ProfHelp-ID logistic regression. [5] ProfHelp-ID classifies *CORRECT* versus other outcomes, combining *INCORRECT* and *HINT_REQUEST* because both indicate that the student lacks the knowledge to answer correctly. ProfHelp-ID predicts whether an attempt will have a *CORRECT* outcome based on general student proficiency, individual differences in proficiency with different attempt types, knowledge component easiness, and a history of prior practice with the knowledge component. Compared to a logistic regression with M parameters, a baseline-category model estimates up to $(K-1)M$ parameters, i.e., twice the number in ProfHelp-ID.

$$logit(\Pr(Y_{kph} = k)) = \boldsymbol{Z_h \lambda_{kp}} + \sum_{j \in KC} (\beta_{kj} + \gamma_{kj} s_{pj} + \rho_{kj} f_{pj})$$

Equation 1: ProfHelp-Multinomial

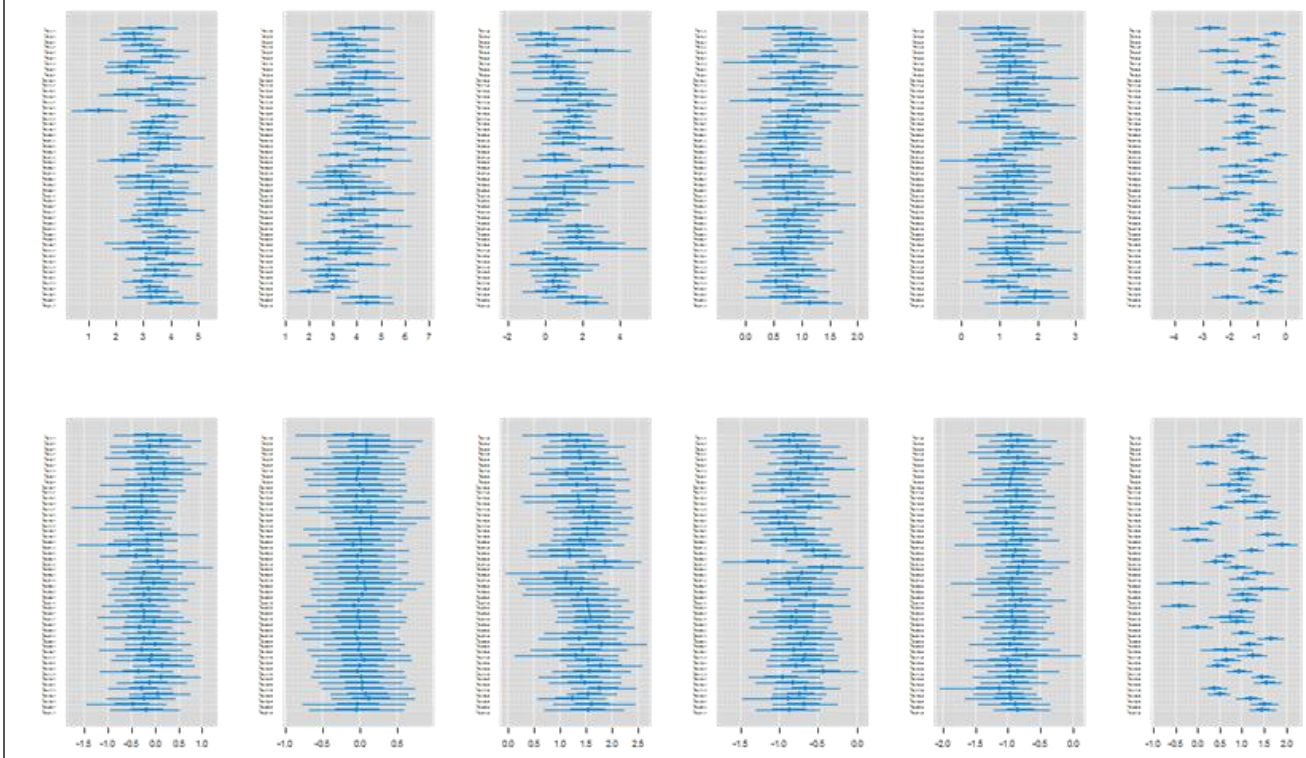

Figure 2: Logit estimates of λ_{kph}. TAH-NRI (top row) and proficiency (bottom). Left to right, the blocks are attempts after feature-pointing hints, principle-stating hints, bottom-out hints, one incorrect, two incorrects, and on first attempts. Each bar shows a 95% Credible Interval for λ_{kph} for one student, and the dot shows the median estimate. The X axes differ across blocks.

$\boldsymbol{\lambda}$ is a matrix, with one parameter for each pupil p, attempt type h and logit comparison k. The 6 possible attempt types are first attempts, attempts directly after each of three types of hints, and attempts directly after a first or second incorrect outcome on the given step. The subscript k implies that the parameter varies across the $K-1$ comparisons. Thus, for pupil p and attempt type h, λ_{1ph} represents TAH-NRI, and λ_{2ph} represents the proficiency, e.g., $\lambda_{1,2,3}$ is TAH-NRI of student 2 on attempts directly following level-2 (principle-stating) hints. As the exploratory data analysis suggests, first-attempt performance may relate to both TAH-NRI and proficiency. Accordingly, Z is a fixed $H \times H$ matrix such that the vector product of $\boldsymbol{Z_h}$ and $\boldsymbol{\lambda_{kp}}$ makes the estimate over first-attempts λ_{kp1} a reference level for estimates for the same k and p on other attempt types (i.e., where $h \neq 1$).

Pupil parameters are partially pooled, $\boldsymbol{\lambda_{kp}} \sim MVN_h(\boldsymbol{\Lambda_k}, \boldsymbol{\Sigma})$. Thus, each student's vector $\boldsymbol{\lambda_{kp}}$ is based on the averages $\boldsymbol{\Lambda_k}$ across all students, and each student's contribution to $\boldsymbol{\Lambda_k}$ is weighted by the number of observations for the student. The hyperprior for $\boldsymbol{\Lambda_h}$ is $N_h(0{,}1000)$, i.e., uninformative. We estimate the variance of the per-pupil parameters (diagonal of matrix $\boldsymbol{\Sigma}$), and impose a structure of zero covariance (off-diagonal cells in $\boldsymbol{\Sigma}$).

Other parameters pertain to knowledge component j. By analogy with per-pupil TAH-NRI and proficiency, β_{1j} represents the attractiveness to hints of KC j, and β_{2j} represents the KC's easiness. The slopes γ and ρ are the effects of student p's prior first-attempt successes s_{pj} and failures f_{pj} on the increased likelihood of a hint (for $k = 1$) and correct response ($k = 2$).

In sum, the model estimates the main effects and individual differences in proficiency on each attempt type, and the main effects and individual differences in TAH-NRI on each attempt type. Unlike the percentages used in exploratory analysis, these estimates account for other predictors, e.g., prior practice, and the model fitting yields credible intervals (CI) about these parameters.

4. RESULTS AND DISCUSSION

The model-fitting indicates substantial individual differences in first-attempt TAH-NRI (Figure 2, top right) and in first-attempt proficiency (bottom right). Estimates of first-attempt TAH-NRI are negative for almost all pupils, with posterior 95% CI for $\Lambda_{1,6}$ ranging -1.71 to -1.14, implying that first attempt hint-requests were unlikely. First-attempt proficiencies are about 1.0, i.e., students in this dataset tend to answer correctly on a first attempt.

Taking first-attempt TAH-NRI for each student as a reference level, attempts after hints are very positively associated with hint-requests rather than incorrects: feature-pointing and principle-stating hints (top left two blocks) are strongly positive for all students, and even bottom-out hints (top row, third block) trend positive for many students.

With first-attempt proficiency as a reference level, proficiencies on attempts after feature-pointing and principle-stating hints (bottom left two blocks) tend to neutral, with (-0.40, 0.08) and (-0.27, 0.25) posterior 95% CIs for $\Lambda_{2,1}$ and $\Lambda_{2,2}$, respectively. In other words, by decoupling incorrect outcomes from hint requests, a multinomial model removed the strongly negative "effect" of these hints estimated by a binary model. [5] Moreover, individual differences in proficiency with hints [5] may be due to differences

in TAH-NRI (top left three plots) rather than differences in proficiency (bottom left three plots). It is plausible that correct responses are more likely after bottom-out hints than other hints, since bottom-out hints often reveal the correct response.

TAH-NRI is unlike other forms of help-seeking, including gaming-the-system behavior that involves "attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly" [3]. Neutral-to-positive proficiency with hints implies that students try to learn from hints, suggesting that TAH-NRI signifies learning, not gaming. Further, harmful gaming behavior was observed in only 8% to 27% of students; we find positive TAH-NRI for almost all students.

TAH-NRI is positive directly after a first incorrect and second incorrect outcomes, with (0.65, 1.05) and (1.15, 1.66) posterior 95% CIs for $\Lambda_{1,4}$ and $\Lambda_{1,5}$, respectively. By contrast, proficiency after a first incorrect or a second incorrect is negative, with 95% CI for $\Lambda_{2,4}$ =(-0.89, -0.64) and for $\Lambda_{2,5}$ = (-1.08, -0.71).

In sum, once we adjust for first-attempt tendency to request hints, we find that students use the hint system extensively, although there are ample individual differences. Students request hints more often after a previous hint request than after an incorrect.

By contrast, once we adjust for first-attempt proficiency, students differ less in proficiency than in TAH-NRI. The "main effect" of hints is neutral (feature-pointing and principle-stating hints) or positive (bottom-out hints). We caution that a better evaluation of these hint types would consider a variety of hint sequences.

4.1 Model Adequacy

As often happens in classification, the model is biased (Table 3) in favor of predicting the majority class (*CORRECT*).

Table 3: Confusion matrix

	Predicted *INCORRECT*	Predicted *HINT_ REQUEST*	Predicted *CORRECT*
INCORRECT	607	951	3123
HINT_ REQUEST	1017	2078	1195
CORRECT	235	953	15778

Attempts after a first or a second incorrect are more likely to be misclassified than indicated by their prevalence in the full dataset (Table 4). This is puzzling. At the heart of ProfHelp-Multinomial is the PFA model of first-attempt performance [7]. Performance after one incorrect or two incorrect outcomes is highly correlated with first-attempt performance (Table 1), so PFA should be reasonably accurate for these attempts, and the λ_{kph} parameters in ProfHelp-Multinomial should further improve accuracy.

For future work, first, in combination with prior findings on help-avoidance and help-abuse, our results imply that statistical models not only take into account local transitions from attempt to attempt, but longer sequences of attempts. This is consistent with the help-seeking model of Aleven et al [1]. Bridging data-mining and theoretical approaches will lead to a model that more accurately reflects student help use.

Second, the fact that students often make multiple attempts to solve a problem-step without hints coupled with ProfHelp-Multinomial's poor predictive accuracy of performance on attempts after incorrects calls for data mining and other research to understand same-step performance after incorrect outcomes.

Table 4: Prevalence of observations *vs.* prediction errors

Attempt Type	Percent of Dataset	Percent of Prediction Errors
First attempts	67	50
After FP Hint	7	10
After PS Hint	4	5
After BOH	6	5
After 1st Incorrect	11	20
After 2nd Incorrect	4	9

5. CONCLUSIONS

This work advances the study of same-step help use in ILE. Students have a tendency to request multiple hints in a row rather than risk an error. Our analysis improves on prior analyses of TAH-NRI [8] in that we find that TAH-NRI may differ based on the type of attempt, and it persists after accounting for proficiency, for a knowledge component's attractiveness to hints, and for prior practice. TAH-NRI is distinct from other help-abuse behavior, e.g., from gaming-the-system. Further, there are persistent individual differences among students in TAH-NRI.

Formalizing TAH-NRI in the ProfHelp-Multinomial model alleviates the selection bias that caused another model to estimate that hints had negative effects. The ProfHelp-Multinomial results suggest that students make a strategic decision to get help and stick with it across attempts, i.e., the decision to try to solve vs. to request a hint is not an independent decision at each attempt.

The improved understanding of help-seeking developed here is a step towards developing effective and efficient ILE, including systems that adapt to individual differences among students.

6. REFERENCES

[1] Aleven, V. et al. 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.

[2] Aleven, V. and Koedinger, K.R. 2001. Investigations into help seeking and learning with a cognitive tutor. *Papers of the AIED-2001 Workshop on help provision and help seeking in interactive learning environments* (2001), 47–58.

[3] Baker, R.S.J. d. et al. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*. 18, 3 (Jan. 2008), 287–314.

[4] Beck, J.E. et al. 2008. Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. *Intelligent Tutoring Systems*. B.P. Woolf et al., eds. Springer. 383–394.

[5] Goldin, I.M. et al. 2012. Learner Differences in Hint Processing. *Proceedings of 5th International Conference on Educational Data Mining* (Chania, Greece, 2012), 73–80.

[6] Goldin, I.M. and Carlson, R. 2013. Learner Differences and Hint Content. *Proceedings of 16th International Conference on Artificial Intelligence in Education* (Memphis, TN, 2013).

[7] Pavlik Jr, P. et al. 2009. Performance Factors Analysis - A New Alternative to Knowledge Tracing. *Proceedings of 14th International Conference on Artificial Intelligence in Education* (Brighton, England, 2009), 531–538.

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