Student Profiling from Tutoring System Log Data: When do Multiple Graphical Representations Matter?

Ryan Carlson
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
ryancarlson@cmu.edu

Konstantin Genin
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA, USA
kgenin@andrew.cmu.edu

Martina Rau
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA
marau@cs.cmu.edu

Richard Scheines
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA, USA
scheines@cmu.edu

ABSTRACT

We analyze log-data generated by an experiment with Fractions Tutor, an intelligent tutoring system. The experiment compares the educational effectiveness of instruction with single and multiple graphical representations. We extract the error-making and hint-seeking behaviors of each student to characterize their learning strategy. Using an expectation-maximization approach, we cluster the students by learning strategy. We find that a) experimental condition and learning outcome are clearly associated b) experimental condition and learning strategy are not, and c) almost all of the association between experimental condition and learning outcome is found among students implementing just one of the learning strategies we identify. This class of students is characterized by relatively high rates of error as well as a marked reluctance to seek help. They also show the greatest educational gains from instruction with multiple rather than single representations. The behaviors that characterize this group illuminate the mechanism underlying the effectiveness of multiple representations and suggest strategies for tailoring instruction to individual students. Our methodology can be implemented in an on-line tutoring system to dynamically tailor individualized instruction.

1. INTRODUCTION

Multiple graphical representations (MGRs) are ubiquitous in math and science instruction: they are frequently used to emphasize complementary conceptual interpretations of complex learning materials. Fraction instruction is one domain in which graphical representations, such as number lines, pie-charts, and rectangles are used to help students overcome the difficulty of the material. Although the educational psychology literature suggests that requiring students to translate between representations supports the creation of deep knowledge structures [6], the experimental results are somewhat ambiguous [1] and the mechanisms underlying these advantages are not well understood [2].

Because student interaction with intelligent tutoring systems (ITSs) generates very fine-grained behavioral and outcome data, these systems are well-suited for conducting experiments on the effect of MGRs on learning outcomes [14]. Machine learning methods can be profitably applied to identify the kinds of students whose learning outcomes are improved by MGRs and the factors mediating their success [19]. Such insights enable developers of ITSs to design individualized instructional support that can make learning with MGRs even more effective. This may involve encouraging students to reflect on the material with self-explanation prompts [17] or detecting ineffective strategies and implementing interventions on-the-fly. Work in the latter area ranges from detecting abuse of the ITS hint system and other "gaming" behaviors [8, 7] to providing spontaneous help to students lacking the metacognitive skills to know when they could use a hint [3, 4, 5].

Prior research conducted on elementary-school students working with a Fractions Tutor suggests that prompting students to self-explain while working with MGRs improves their educational effectiveness [17]. Subsequent studies examining error-rate, hint-use and time-spent in tutor's log failed to identify variables that mediate the effectiveness of MGRs [16]. The mechanisms by which multiple graphical representations improve learning outcomes remain poorly understood.

We conjecture that previous efforts to identify mediating factors were frustrated by heterogeneity in the problem-solving habits and behaviors of the student population under investigation. Using a mixture modeling technique, we cluster

students by the patterns of interaction with the tutor in the log-data that characterize their learning strategy. Clustering based on student characteristics has proved successful in grouping students into meaningful subpopulations across both collaborative [15] and individual [9, 13, 20] educational environments.

Four strategic profiles emerge from our analysis, each with a natural interpretation. Two of the profiles are characterized by a low propensity to seek help from the tutor. In one of these the students are simply confident: they make few errors, solicit little help and don't seem to need any. In the other the students are reluctant to solicit help even though they seem like they need it: they make a relatively large number of mistakes but make little use of the support mechanisms the tutor provides. We characterize this second class as "stubborn" without intending any pejorative connotations. A third class is highly interactive: they make many mistakes, seek assistance readily and frequently exhaust the hints available in a given problem. Students in the fourth class occupy a middle ground between the interactive and the stubborn: they make an average number of mistakes and will eventually seek help when they are having trouble.

We proceed to explore how the experimental conditions affect post-test outcomes. Confirming previous results [16], we find that students in the multiple-representation condition had greater learning gains than those in the single-representation condition. MGRs seem to have a robust and positive effect on long-term knowledge consolidation. We then explore the effect of multiple representations in the sub-populations defined by each strategic profile. We first establish independence between learning strategy and experimental condition. This suggests that we are detecting pre-existing strategic profiles, rather than artifacts of the experimental setup. Most interestingly, we discover that learning gains from MGRs depend heavily on learning strategy. Students exhibiting a "stubborn" profile profited substantially from instruction with multiple rather than single representations. For the remaining students, experimental condition and learning gain were independent. We conjecture that "stubborn" students lack the metacognitive skills to judge when their learning strategies are failing. These students are the most sensitive to pedagogical decisions because they are the least equipped to structure and manage their own learning.

Section 2 of what follows describes the initial experiment and elaborates on the differences between the representational conditions. We describe our feature extraction process and modeling decisions in Section 3. Section 4 summarizes the results of the model estimation and statistical analysis of the effects of multiple representations at the population and sub-population levels. We suggest profitable future directions in Section 5.

2. EXPERIMENT

In the Spring of 2010, Rau conducted an experiment wherein 290 4^{th} and 5^{th} grade students worked with a Fractions Tutor for about 5 hours of their mathematics instruction. Students were randomly assigned to one of five experimental conditions, which varied by the frequency with which students would be presented with a new fraction representation (see Figure 1). Students in the SINGLE representation condition worked exclusively with either a number line, a circle or a rectangle. Students in the FULLY INTERLEAVED condition saw a different representation than was used in the preceding problem. Students in the intermediate conditions went longer before seeing a different representation.

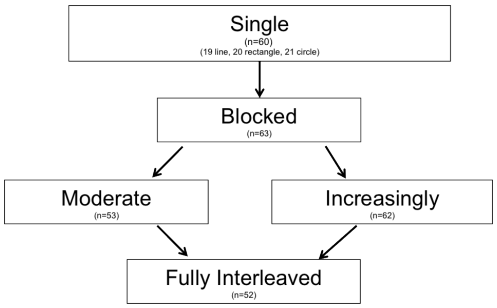

Figure 1: A partial ordering of experimental conditions by the frequency with which a new representation is presented.

When interacting with different graphical representations of fractions, students were able to drag-and-drop slices of a pie chart, for example, into separate areas. They were also able to experiment with changing the number of subdivisions in each graphical representation. Students received a pre-test on the day before they began working with the tutor and an immediate post-test on the day after they finished. Students also took a delayed post-test a week after the first. Previous investigation found that students in the MGR conditions significantly outperformed students in the single representation condition on the delayed post-test [16, 18].

3. METHOD

We proceed in three stages: (1) we extract features characterizing error and hint-seeking behavior from the data logs, (2) we transform the longitudinal log data into a cross-sectional form, with one observation per student, and (3) we estimate a mixture model to identify sub-populations of students, using AIC and BIC to select the number of classes.

Once we have clustered our students by their learning strategy, we investigate the interaction between the strategies and the experimental conditions. We construct a contingency table binning the experimental conditions into the clusters estimated by the mixture model. We then run a Chi-squared test for independence between experimental condition and learning strategy. Chi-squared tests are also run to investigate dependence between pre-test outcome and strategy, strategy and post-test outcome and the conditional dependence of outcome and experimental condition, given a strategic profile.

3.1 Extracting Features

The Fractions Tutor captures a detailed log of each student's interactions with the tutor. It stores a time series of correct and incorrect answers, hint requests, interface selections and durations between interactions. Previous analysis [16] extracted the average number of errors made per step, the average number of hints requested per step, and the

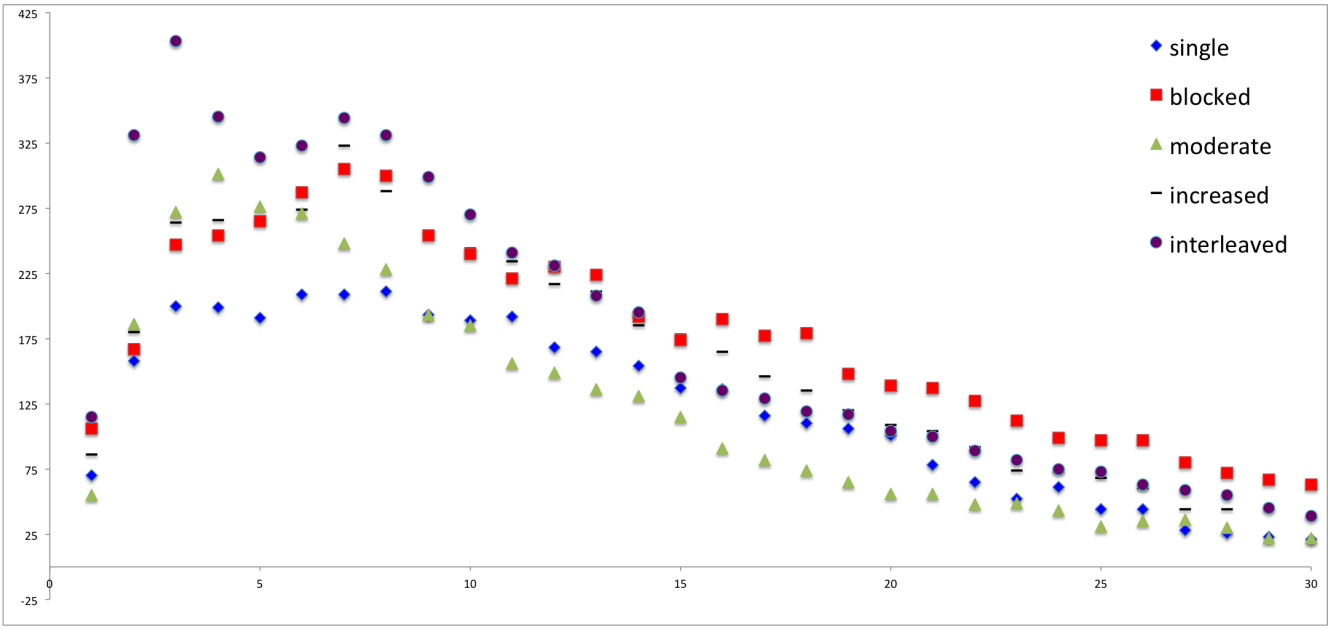

Figure 2: The x-axis represents the n_{th} interaction with the tutor across all problems. The y-axis is the total number of hints requested at the n_{th} step.

average time spent per step from the log data. These variables were used to characterize gross behavioral strategies and dispositions. Similarly, we include the average number of hints requested (HintsRequested) and number of errors (NumErrors) made per *problem* by each student. We also extract the average number of bottom-out hints (NumBOH) per student per problem – this is the average number of times a student exhausts the available hints in a given problem. We also note that it is not always the average of these features that best characterizes a student. For example, examination of the distribution of hints requested per step across experimental condition, shows a telling picture.

Note that students who received only one representation start out requesting the fewest hints, but students in the moderate condition eventually need fewer (see Figure 2). Also, students in the interleaved condition tend to request many hints in the early steps of a problem, potentially reflecting the cognitive load associated with translating between representations [1]. Such considerations suggested that exploiting the timing of student interactions within a problem might expose structural features obscured by stepwise averages (as used in [16]). We fit geometric distributions to the number of steps taken before the first hint request (FirstHintGeometric) and to the number of errors before the first hint (StubbornGeometric). The estimated parameter is used to characterize the student's hint-seeking propensity in general and hint-seeking propensity when faced with adversity. For example, students in the first quintile of StubbornGeometric seek help soon after making a mistake, whereas students in the fifth quintile don't change their hint-seeking behavior even after making a large number of errors. Students in the first quintile of FirstHintGeometric are likely to request hints early in a problem, whereas students in the fifth quintile are unlikely to request hints at any point.

3.2 Expectation-Maximization Clustering

Expectation-Maximization (EM) clustering is a modeling technique that determines subtypes based on multinomial distributions. We use the model to categorize students into subpopulations using discretized versions of the features described above. Table 1 shows summary statistics and cut-off points for the extracted features. The model maps a set of observed categorical variables onto a set of inferred classes.

We note that the categorical nature of the model has the potential to add some noise, since we must select numeric cutoffs to transform our variables into nominals. However, categorical models can offer greater interpretability by allowing us to organize our data into a small set of variables, which forms the basis for categorizing students into a small set of meaningful, homogenous groups. Furthermore, it is not unreasonable to suspect that our variables are in some sense "truly" categorical [10, pp8–9]. EM clustering requires a relatively small set of variables to train the model. As the number of training variables increases, the number of model parameters blows up and the model becomes overspecified.

Unlike some common clustering algorithms (e.g., k-means), EM produces "fuzzy" clusters (i.e., probability distributions over features for each class). We use these probability distributions in our qualitative discussion about the subpopulations (Section 4.1), however we ultimately need to identify each student's most likely class. For each student s and class c we calculate

$$\arg\max_{c} P(S = s \mid C = c) \tag{1}$$

where the probabilities are determined by the EM algorithm.

Table 1: Summary Statistics for Variables Used in Clustering

	mean	sd	median	min	max	20%	40%	60%	80%	100%
HintsRequested	0.78	1.27	0.34	0	11.22	0.06	0.19	0.5	1.31	11.22
NumErrors	2.21	1.27	1.92	0.34	8.39	1.15	1.7	2.18	3.19	8.39
FirstHintGeometric	0.35	0.27	0.27	0.04	1	0.13	0.2	0.33	0.57	1
Stubborn Geometric	0.36	0.21	0.31	0.07	1	0.19	0.27	0.38	0.47	1
NumBOH	0.04	0.08	0	0	0.62	0	0	0.01	0.05	0.63

We still need to fix N, the number of classes. We use two complexity-penalized log-likelihood scores to select an appropriate N: Akaike information criterion (AIC) and Bayesian information criterion (BIC). Plotting these statistics as we increment the number of classes, we look for a "knee" where both statistics either bottom-out or level off to identify the optimal value of N. To run analysis, we used `poLCA`, a freely available R package.[1]

4. RESULTS

In the sections that follow we analyze the results of our clustering algorithm. We describe the strategic profiles that were generated and characterize the students fitting each profile. We then consider the relationships between our variables of interest: *(a)* adjusted delayed post-test score, *(b)* experimental condition, and *(c)* learning strategy. Specifically, we run a series of Chi-squared tests for independence to determine how each variable relates with the others, commenting on the importance of each comparison. Finally, we explore the stability of these classes, which bears on whether future systems could detect students' strategic profiles in real time.

4.1 Exploring the Learning Strategies

Figure 3 shows the parameter selection process described in Section 3.2. Note that we chose to model four classes because BIC bottoms out and AIC levels off at that point.

After selecting the appropriate N parameter, we extract membership probabilities for the individual students. Given a strategic profile, we can estimate the probability distribution over each feature, and use Equation 1 to identify the most likely profile for each student.

The feature distributions over each profile are represented graphically in Figure 4. Each feature is listed along the horizontal x-axis, the value each variable takes is along the front-to-back y-axis, and the probability that the feature takes that value is given along the vertical z-axis. For example, consider the HintsRequested feature (average hints requested per problem) in the "interactive" class. In that class, with high probability, students requested many hints (i.e., the highest categorical value for hints) per problem on average. As another example, students in the "moderate" class are more likely to make a moderate number of errors,

[1] `http://userwww.service.emory.edu/~dlinzer/poLCA/`

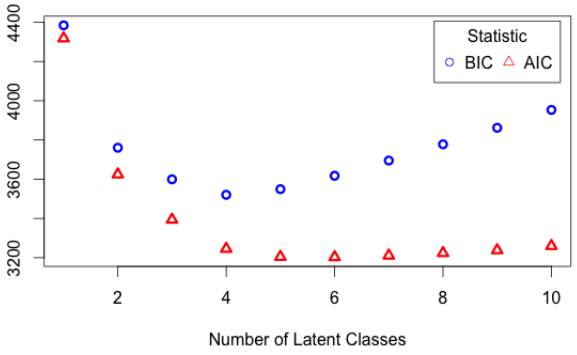

Figure 3: AIC and BIC over increasing number of clusters. BIC bottoms out and AIC levels off at four clusters, so we conclude that four clusters best fits the data.

though other error levels also occur with nontrivial probabilities. Lower values of FirstHintGeometric and StubbornGeometric indicate a steep geometric slope, corresponding to a higher hint-seeking propensity and stubbornness, respectively.

How do we interpret cluster membership? Students in Class 1 are "Moderate", they ask for a moderate number of hints, make a moderate number of errors, and are moderately responsive to the interface. Students in Class 2 are "Interactive", they make a lot of errors, but respond by requesting many hints. These students are proactive in asking for help and are not shy about using the resources the Fractions Tutor makes available. Students in Class 3 are "Confident", they don't ask for hints, but they don't seem to need them (since they make few errors). Finally, we call students in Class 4 "Stubborn" because they are fairly mixed in error-profile but they don't respond to mistakes with hint-requests. These students are not using all the resources that the Fractions Tutor makes available.

4.2 Condition and Outcome

We use normalized learning gain at the delayed post-test as our measure of student improvement.

$$\text{Learning Gain} = \frac{\text{DelayedPostTest} - \text{PreTest}}{1 - \text{PreTest}}$$

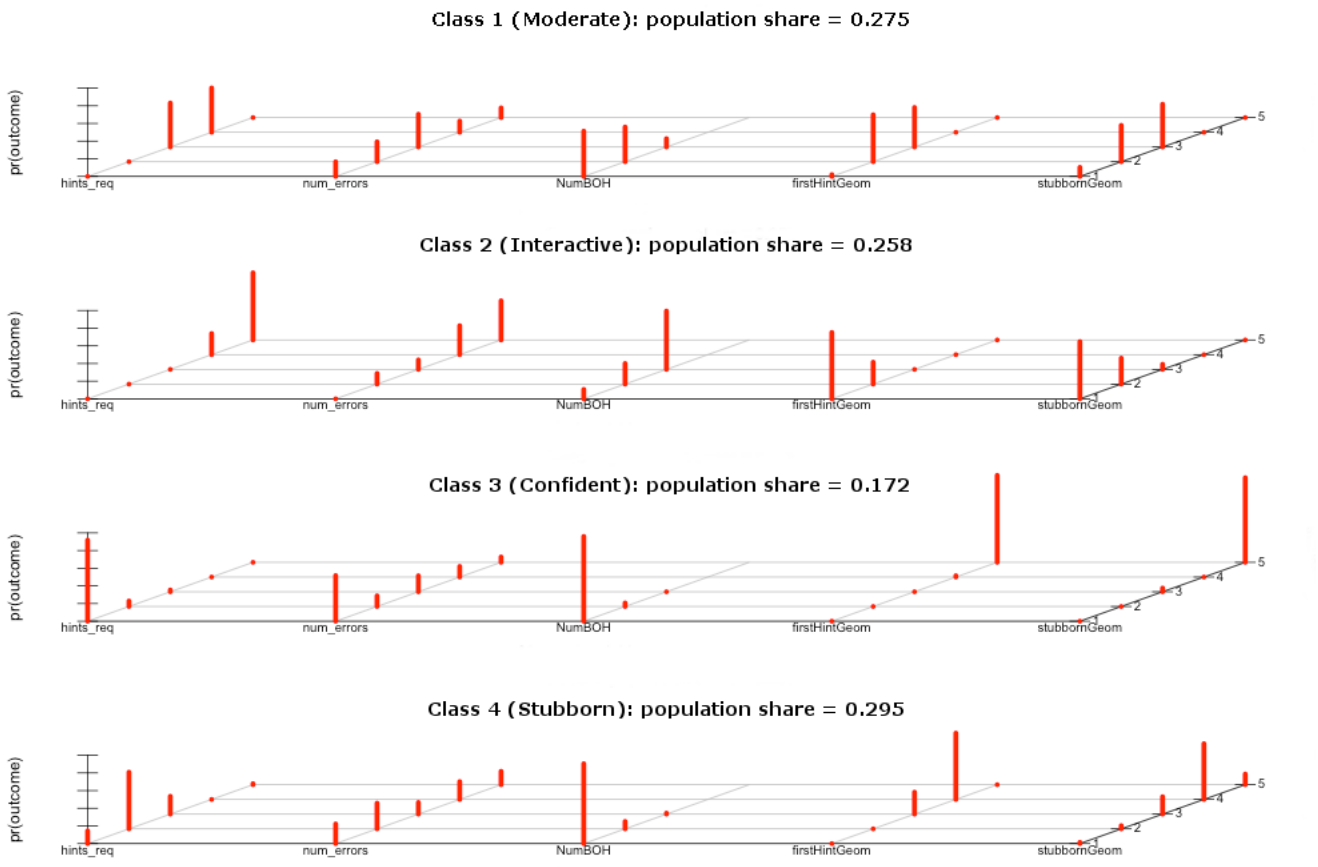

Figure 4: Visualization of feature distributions for each learning profile. The left-to-right x-axis identifies each feature, the front-to-back y-axis identifies which value that feature takes, and the top-to-bottom z-axis describes the probability that the feature takes the value. Thus, given a feature and a class, the z-axis also describes the probability distribution over that feature in that class.

We then construct terciles of the Adjusted Delayed Post-Test Score and run a Chi-squared test for independence of outcome from experimental condition. Confirming previous results, we reject independence at a p-value of .024 (see Table 2). As expected, students in the multiple representation conditions were more likely to be in the second or third tercile of adjusted delayed post-test score, whereas students in the single representation condition were more likely to be in the first.

4.3 Learning Strategy and Test Scores

We would expect that a student's learning strategy would predict (and perhaps cause) their ultimate educational outcome. To test this intuition, we calculate a Chi-squared statistic for independence of learning strategy from normalized delayed post-test gain. We reject independence at a p-value of .0075 (see Table 3). The behaviors encoded by strategic profile seem highly relevant to knowledge consolidation in the long run. Students in the moderate class are found mostly in the second and third tercile. These students are implementing a subtle but effective strategy. Their moderation in hint-seeking indicates a level of self-reflectiveness

	33%	66%	99%
blocked	14	29	20
increased	22	20	20
interleaved	13	21	18
moderate	18	13	22
single	30	13	17

$\mathcal{X}^2 = 17.65$, df $= 8$, p-value $=$ **0.024**

Table 2: Experimental Condition by Tercile of Adjusted Delayed Post-Test Score.

that we would expect from students with highly developed metacognitive skills. Students in the interactive class are characterized by a high number of errors, so we are not surprised to find them represented mostly in the first and second terciles. These students are the most likely to exhaust all the hints available in a given problem. If one were looking for students engaging in "gaming" behavior this would be the class to search, perhaps using techniques from [7]. As one would expect, the confident students are likely to end up in the third tercile. The stubborn students are clustered at the extremes: they are more likely to end up in the first or third tercile than the second.

	Learning Gain			Pre-Test		
	33%	66%	99%	33%	66%	99%
moderate	20	35	29	30	32	22
interactive	33	26	14	37	27	9
confident	13	15	22	5	14	31
stubborn	31	20	32	26	22	35

Learning Gain: $\mathcal{X}^2 = 17.52$, df = 6, p-value = **0.0075**
Pre-Test: $\mathcal{X}^2 = 42.3764$, df = 6, p-value = **<0.001**

Table 3: Learning Strategy by Tercile of Normalized Delayed Post-Test and Pre-Test Score

Although we implicitly account for the pre-test scores in our learning gain metric, we also investigate the relationship between learning strategy and pre-test scores (Table 3). As expected, we reject independence between strategic profile and pre-test score, suggesting that these profiles are genuinely meaningful descriptions of student behavior.

Although pre-test score and strategic profile are dependent, the average pre-test score for the "stubborn" students does not differ significantly from the rest of the population.[2] Pairwise t-tests between the four profiles show significant differences in mean pre-test score for all pairs except stubborn and moderate. This analysis suggests that the dependence we detect between experimental condition and outcome for the "stubborn" students does not hinge essentially on pre-test score. If pre-test is an accurate proxy for preparedness, the stubborn students do not occupy a preparedness "sweet-spot" that makes multiple representations uniquely effective. Rather, it seems to be their unique strategic profile that accounts for the effectiveness of MGRs.

4.4 Condition and Learning Strategy

We may also worry that experimental condition is inducing learning strategy. If this were the case, we would suspect that we were picking up on artifacts of the experimental design rather than pre-existing student profiles. However, using the Chi-squared test, condition and cluster membership appear independent (see Table 4).[3] To anticipate Simpson's paradox-type worries, we collapse all four multiple representation conditions (blocked, moderate, increased, interleaved) into a single "multiple representation" condition, but still find independence.[4] These results suggest that our method is detecting genuine student profiles, independent of experimental condition.

	mod.	inter.	conf.	stub.
blocked	13	15	10	25
increased	21	16	10	15
interleaved	17	18	7	10
moderate	18	10	12	13
single	15	14	11	20

$\mathcal{X}^2 = 12.85$, df = 12, p-value = 0.38

Table 4: Experimental Condition by Learning Strategy

4.5 Condition, Outcome and Strategy

Finally, we explore the relationship between learning outcome and experimental condition for each of the strategic profiles we have identified. Interestingly, we find that experimental condition has a substantial effect on learning outcome among the "stubborn" students, but virtually no effect on learning among the "moderate", "interactive", and "confident" (see Table 5). Most students perform in the second and third tercile when given multiple graphical representations, but are overwhelmingly in the first tercile when given a single representation.

Students in the other three classes are not significantly affected by their representation condition. The learning strategies that these students implement seem to make them resilient to representational choice, at least in this experimental regime. Recall that students exhibiting the "stubborn" profile rarely requested hints, even when they encountered difficulty. We speculate that they lack the metacognitive skills to judge when their learning strategies are failing, and thus are not seeking help at appropriate times [4]. They are the most sensitive to pedagogical decisions because they are the least equipped to structure and manage their own learning.

An ITS ought to ensure that these students are targeted with multiple representations, and perhaps other forms of metacognitive support. While not all "stubborn" students improve when given MGRs, the vast majority of them do. An ITS might help scaffold effective learning behaviors by spontaneously offering hints to these students when they appear to need them the most. A teacher informed that a student exhibits this learning profile may try to encourage the student to ask for help and target their metacognitive skills more generally. Moreover, studying this subpopulation seems to be a promising avenue for illuminating the mechanism by which MGRs improve learning outcomes. Future experiments could test the effect of offering spontaneous hint-support to students that fit the "stubborn" profile.

[2]Student's T-test: t = 0.9978, df = 139.602, p-value = 0.3201
[3]We fail to reject independence at a p-value of .38.
[4]$\mathcal{X}^2 = 1.1517$, df = 3, p-value = 0.7646

moderate	33%	66%	99%
blocked	2	8	3
increased	4	9	8
interleaved	4	9	4
moderate	4	5	9
single	6	4	5

$\mathcal{X}^2 = 8.08$, df = 8, p-value = 0.43

interactive	33%	66%	99%
blocked	7	6	2
increased	9	5	2
interleaved	5	8	5
moderate	7	2	1
single	5	5	4

$\mathcal{X}^2 = 6.95$, df = 8, p-value = 0.54

confident	33%	66%	99%
blocked	0	5	5
increased	3	3	4
interleaved	2	2	3
moderate	3	4	5
single	5	1	5

$\mathcal{X}^2 = 7.41$, df = 8, p-value = 0.49

stubborn	33%	66%	99%
blocked	5	10	10
increased	6	3	6
interleaved	2	2	6
moderate	4	2	7
single	14	3	3

$\mathcal{X}^2 = 17.4837$, df = 8, p-value = **0.025**

Table 5: Condition and Tercile of Adjusted Delayed Post-Test Score, by Learning Strategy

We note that there are competing interpretations of our results that also suggest interesting future experiments. Studies have found that well-designed feedback from errors may be very effective for improving learning outcomes [12]. It may be that "stubborn" students, by not shying away from mistakes, are taking advantage of a more effective support system than students who avoid mistakes by soliciting hints. Since instruction with multiple representations is generally more difficult, stubborn students in a multiple representation condition would get more of this kind of feedback on average. This interpretation would predict that students in the "interactive" profile would benefit if some hints were withheld [11]. However, this hypothesis could only be tested by subsequent experiments.

4.6 Profile Stability

If an intelligent tutoring system could implement our classification methodology on-the-fly, it could tailor its pedagogical interventions to the needs of the individual student. To substantiate the promise of the methodology we investigate how efficiently the algorithm stabilizes to the final classification. To measure this, we first cluster on the entire corpus and assign each student to their most likely profile. We then artificially subset the data by restricting the number of problems seen by the clustering algorithm, compute the proportion of students who are in their "final" profile, and then iteratively increase the size of the subset. This simulates how well our algorithm identifies student profiles as they make their way through the material.

Figure 5 shows the percentage of total data used to estimate the model plotted against the proportion of students assigned to their final strategic profile. At each iteration, we look at an additional 10 problems from each student and re-estimate the cluster assignments. The regression estimates that 63% of the data is sufficient to classify three quarters of

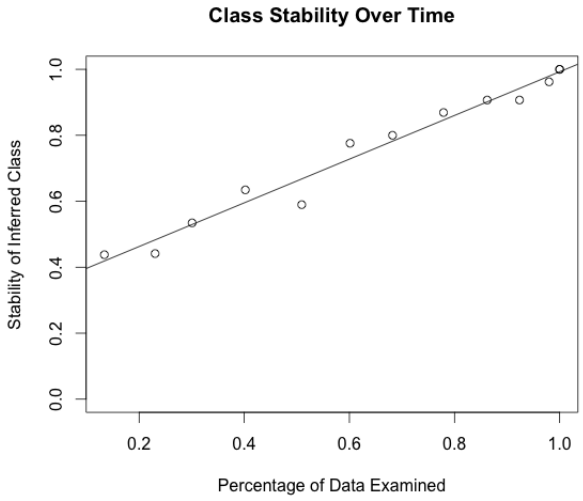

Figure 5: We measure the number of students who were classified into their ultimate strategic profile as the amount of data available to EM is increased. We see that at about 60% of the data we can classify about 75% of the students into their ultimate profile.

the students into their ultimate cluster. Thus, after seeing about 60 problems – about two days of classroom instruction – a dynamic intelligent tutoring system might intervene on students who fit the "stubborn" profile by ensuring that they are presented with multiple graphical representations, offering them spontaneous hints or targeting them with some other form of metacognitive support.

5. CONCLUSION & FUTURE WORK

We estimated an expectation maximization clustering model to classify students into four strategic profiles based on their error-rates and hint-seeking behaviors. We detected an effect of experimental condition on post-test outcome only in the class of students characterized by high-error rate and low hint-seeking propensity. That is, students who did not seem to take full advantage of the resources that the Fractions Tutor offered were the ones most strongly affected by experimental condition. These students may not have the metacognitive skills required to know when to seek hints.

Our methods could be used by ITS designers to detect students with this profile in real time. Tutoring systems could then intervene to target these students with MGRs, scaffold their hint-seeking behaviors or target them with other forms of metacognitive support. Future research into the mediating mechanisms of multiple representations could leverage our results to identify the relevant student sub-populations to investigate. Our post-hoc analysis is not designed to identify the cognitive processes underlying the student's problem solving behavior, so interviews or a cognitive task analysis with students who fit the "stubborn" profile could reveal more details about their experience than we can detect from the log data. Additional investigation into different features may help further characterize student behavior and could help us more accurately group students into relevant subpopulations. Although our analysis seems to have revealed interesting differences in student learning strategies, more informative features constructed from log data may do better. Constructing more informative features, for example, might allow us to separate the "stubborn" students into those who did and did not benefit from multiple graphical representations.

6. REFERENCES

[1] S. E. Ainsworth. The functions of multiple representations. *Computers and Education*, 33(2-3):131–152, 1999.

[2] S. E. Ainsworth. Deft: A conceptual framework for learning with multiple representations. *Learning and Instruction*, 16(3):183–198, 2006.

[3] V. Aleven and K. R. Koedinger. Limitations of student control: Do students know when they need help? In G. Gauthier, C. Frasson, and K. VanLehn, editors, *Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000*, pages 292–303, 2000.

[4] V. Aleven, B. McLaren, I. Roll, and K. R. Koedinger. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, 2006.

[5] V. Aleven, E. Stahl, S. Schworn, F. Fischer, and R. M. Wallace. Help seeking and help design in interactive learning environments. *Review of Educational Research*, 73(2):277–320, 2003.

[6] S. A. Ambrose, M. W. Bridges, M. DiPietro, M. C. Lovett, and M. K. Norman. *How Learning Works*. Jossey-Bass, 2010.

[7] R. Baker. Differences between intelligent tutor lessons, and the choice to go off-task. In *Proceedings of the 2nd International Conference on Educational Data Mining*, 2009.

[8] R. Baker, A. Corbett, I. Roll, and K. R. Koedinger. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 2009.

[9] C. R. Beal, L. Qu, and H. Lee. Mathematics motivation and achievement as predictors of high school students' guessing and help-seeking with instructional software. *Journal of Computer Assisted Learning*, 24:507–514, 2008.

[10] L. M. Collins and S. T. Lanza. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Wiley Publishing, 2009.

[11] K. R. Koedinger and V. Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. *Education Psychology Review*, 19:239–264, July 2007.

[12] J. E. McKendree. Effective feedback content for tutoring complex skills. *Human Computer Interaction*, 5:381–414, 1990.

[13] A. Merceron and K. Yacef. Clustering students to help evaluate learning. *Technology Enhanced Learning*, pages 31–42, 2005.

[14] A. Newell and P. Rosenbloom. Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and Their Acquisition*, 1981.

[15] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaiane. Clustering and sequential pattern mining of online collaborative learning data. *Knowledge and Data Engineering, IEEE Transactions*, 21(6):759–772, 2009.

[16] M. Rau and R. Scheines. Searching for variables and models to investigate mediators of learning from multiple representations. In *International Conference on Educational Data Mining*, 2012.

[17] M. A. Rau, V. Aleven, and N. Rummel. Intelligent tutoring systems with multiple representations and self-explanation prompts support learning of fractions. *International Conference of Artificial Intelligence in Education*, pages 441–448, 2009.

[18] M. A. Rau, V. Aleven, Z. Tunc-Pekkan, L. Pacilio, and N. Rummel. How to schedule multiple graphical representations? a classroom experiment with an intelligent tutoring system for fractions. In *Proceedings of International Conference of the Learning Sciences*, 2012.

[19] J. Self. The application of machine learning to student modelling. 14(3–4):327–338, 1986.

[20] C. Unsupervised and S. C. to Build User Models for Exploratory Learning Environments. Saleema amershi and cristina conati. *Journal of Educational Data Mining*, 1, 2009.