# Modeling Student Benefit from Illustrations and Graphs

Michael Lipschultz and Diane Litman

Computer Science Department, University of Pittsburgh
{lipschultz,litman}@cs.pitt.edu

**Abstract.** We examine a corpus of physics tutorial dialogues between a computer tutor and students. Either graphs or illustrations were displayed during the dialogues. In this work, stepwise linear regression, augmented to remove unwanted terms, is used to build models that identify situations when each graphic may aid learning. Our experimental results show that grouping students by pretest score, then by gender produces a model that significantly outperforms the baseline.

**Keywords:** student modeling, ITS, dialogue, graphs, illustrations, physics.

## 1 Introduction

One-on-one tutoring between a student and a human tutor is a very effective method of instruction [10]. Intelligent tutoring systems (ITS) have been developed to provide one-on-one tutoring, but from a computer-based tutor rather than a human tutor, and have been shown to improve student knowledge [17].

Visual representations, such as illustrations and graphs, are one method used to convey information to students thought to help them learn concepts. Illustrations use images to represent concepts [15,9], whereas graphs convey concepts primarily through such graphs as bar graphs or line graphs [15]. While much of the ITS research has made the assumption that one representation is best for everyone, differences exist between representations. Illustrations are easier for novices to interpret [12], but have surface features that may distract students [8]. Graphs can help students connect descriptions of situations to the base concepts [16], but students are more likely to make mistakes with them [14]. Researchers have thus examined the benefits of using multiple representations. Helping students become fluent in multiple representations and to be able to translate between them are beneficial [15]. Research into using multiple representations during tutoring tends to treat all students as identical; the switching of representations are on a fixed schedule [13,15]. However, research suggests that there are differences among students, leading to some representations being more beneficial than others. Student differences to consider include gender [14], spatial reasoning ability [9], and skill with domain concepts [9]. Adapting to students in other instances have had success, such as uncertainty and motivation leading to increased persistence and better learning gains [1,6].

S. Trausan-Matu et al. (Eds.): ITS 2014, LNCS 8474, pp. 436–441, 2014.

This paper explores building models to predict when illustrations and graphs benefit learning. We first describe an algorithm that constructs such models using stepwise linear regression augmented to conform to certain syntactic constraints. We then examine the models learned and find that models including both pretest score and gender when describing tutoring situations perform best.

## 2 Corpus

The data comes from a study comparing the effectiveness of showing illustrations versus graphs during conceptual physics tutoring with an ITS [11]. Subjects solved a physics problem in Andes [17], with the Rimac physics coach walking them through problem solving [7]. Andes presented the problem statement and a visual representing the situation described. Rimac provided instruction on solving the problem through a typed natural language dialogue. After solving the problem, subjects engaged in a reflection dialogue designed for students to reflect on concepts; it was a typed natural language discussion with a computer tutor. It began with a question on a key concept from the problem and after answering this question, the student has a discussion of the concept with the tutor. During this discussion, visuals were shown to help explain concepts.

Subjects saw only illustrations or only graphs during tutoring; the visuals presented the same information. Problems, reflection questions, and their orders, remained the same. Twenty-nine college students without college-level physics were recruited and randomly assigned one of the visuals to see. They began by filling out a background survey then completed a standard test for determining spatial reasoning ability [5]. They took a pretest to measure their incoming physics knowledge. At the end of tutoring, they took an isomorphic, counterbalanced post-test. We have 2043 data points at the utterance level.

Prior work on this corpus found differences from the pooled data using ANCOVAs [11]. This paper presents work on mining the data to learn models that can identify situations *when* illustrations or graphs were beneficial for learning.

### 2.1 Features

Features similar to those below have been used in previous work on tutoring systems [4,2,3] and have been found useful by cognitive science research on visual representations [14,9]. From this literature, we selected the features we could extract from the data collected during the study.

**Gender** – **Female** or **Male**
**SpatialReason** – score on the spatial reasoning test (**high**, **low**)[1]
**Condition** – experimental condition (**graph**, **illus**tration)
**PreScore** – score on pretest (**high**, **low**)
**WalkThruPctCorrect** – percent of correct answers in the current problem's walk through dialogue with the physics coach (**high**, **low**)

[1] Median splits were performed for ease of interpreting results.

**RQPctCorrect** – percent of correct answers in the current problem's prior reflection dialogue (**high**, **low**)
**ProblemPctCorrect** – percent of correct answers in current problem (both walk through dialogue and prior reflection dialogue(s)) (**high**, **low**)
**SessionPctCorrect** – percent of correct answers in session (**high**, **low**)
**PctThruProblem** – for each problem, how far through the dialogues (walk through and reflection) the subject has gone (**early**, **late**)
**PctThruSession** – how far through tutoring (# completed dialogues) (**early**, **late**)
**KCusage** – whether Knowledge Components (KCs) must be **stated** or **applied**
**ItemDifficulty** – whether the question is **easy** or **hard**, as determined by percent correct on a small pilot study using these dialogues

## 3 Modeling

To build an adaptive policy, we use stepwise linear regression to learn a model that explains the variance in post-test score using interactions between the features above. Standard stepwise regression produces rules that may be contradictory or non-adaptive, which are not helpful in creating an adaptive policy. We augment stepwise regression to address these additional constraints. We also constrain the syntax of the models to better describe the tutoring situation. Thus, we are trying to optimize $r^2$, subject to certain constraints.

The algorithm below shows how to learn an adaptive policy. Once learned, the policy can be applied at every decision point by starting at the top of the list and applying the first that applies.

1. Convert each feature into binary factors, one factor for each feature value. Each factor has a value of either 1 or 0, depending on whether the feature has that particular value for that data point.
2. Run stepwise linear regression on the data subject to syntactic constraints
   **Model** – Models have the form $postscore = \sum \text{terms} + prescore$. Both *postscore* and *prescore* are continuous variables. *Prescore* is included because pretest scores are often correlated with posttest scores; in this corpus it is a trend.
   **Terms** – Create terms by multiplying two or more factors together. Each term contains one Condition factor so that the final model learned can indicate situations when a visual helped or hindered learning. Additional factors in the term describe the situation.
3. Identify problematic term pairs. Problematic terms can be identified by:
   **Contradictory pair** – Two terms with opposite conditions and the other factors are identical. For example, 0.123*ConditionIsGraph* GenderIsFemale and 0.789*ConditionIsIllus*GenderIsFemale contradict each other because the first says to show graphs to females, while the second says to show illustrations.
   **Non-adaptive pair** – Two terms with the same factors, except one is opposite between the two terms. For example, 0.456*ConditionIsGraph* PctThruSessionIsLate and 0.123*ConditionIsGraph*PctThruSessionIsEarly are not adaptive since they say to show graphs regardless of the percent through tutoring.
4. For each problematic term pair, remove the one with the lower absolute value of the coefficient (avc)[2].

[2] We also explored removing both terms, but found that the final models did not perform as well.

**Table 1.** Models are compared across 10-fold cross validation according to adjusted $r^2$ values and their 95% confidence intervals. Italicized rows indicate results significantly better than baseline ($p < 0.05$). Underlined indicates the best result.

| Model | | Adj. $r^2$ | 95% CI |
|---|---|---|---|
| Baseline (Illustration) | | 0.1127 | (0.0896, 0.1358) |
| 1 Factor | | 0.0955 | (0.0737, 0.1172) |
| 2 Factors | *Gender* | *0.1788* | *(0.1428, 0.2148)* |
| | SpatialReason | 0.1488 | (0.1149, 0.1826) |
| | *PreScore* | *0.3499* | *(0.3266, 0.3732)* |
| | PctThruProblem | 0.1007 | (0.0635, 0.1378) |
| | PctThruSession | 0.1180 | (0.0851, 0.1509) |
| 3 Factors (PreScore and ...) | *Gender* | *0.4571* | *(0.4220, 0.4922)* |
| | *SpatialReason* | *0.2817* | *(0.2367, 0.3267)* |
| | *PctThruProblem* | *0.3418* | *(0.3183, 0.3653)* |
| | *PctThruSession* | *0.3087* | *(0.2782, 0.3392)* |

5. With the remaining terms, run multiple linear regression to learn the final model since the coefficient signs may change from the original model.
6. Convert the terms into rules and rank them using avc. The Condition factor indicates the visual to show and the other factors indicate the situation. For negative coefficients, use the visual opposite the one indicated by the Condition factor. Negative coefficients suggest that the visual is detrimental to learning in that situation.

## 4 Results

The models are compared to a baseline, which always predicts showing the same kind of graphic. We choose illustrations since they showed better learning gains. Models are each evaluated using ten-fold cross validation and are compared according to the adjusted $r^2$ value. The performance of the baseline can be seen in the first row of Table 1.

The "1 Factor" model contains only one factor describing the situation, plus the interaction feature Condition. As seen in Table 1, it is not significantly different than the baseline. Since all terms in this model consist of one non-Condition factor, the model can only identify situations by one feature (e.g. GenderIsFemale or PctThruSessionIsLate). This may not be enough to adequately describe situations when illustrations or graphs are more beneficial than the other; the descriptions may be too coarse-grained.

Finer-grained situation descriptions are created by adding more factors to each term. Five features were selected based on prior work suggesting a change in these features can cause large changes in models [11,9]: Gender, SpatialReason, PreScore, PctThruProblem, and PctThruSession. Five "2 Factor" models were created, one for each feature; two perform significantly better than baseline: Gender and PreScore, with PreScore significantly better than other models seen so far. Thus, keeping PreScore as the second feature, we add a third factor to this model, drawing from the same set of features. As seen in Table 1, all four "3 Factor" models

**Table 2.** Rules for the best 3 Factor model: PreScore*Gender

| Female High Pretesters (n = 8) | Female Low Pretesters (n = 9) |
|---|---|
| 1. If WalkThruPctCorrect=Low, show Graph<br>2. If RQPctCorrect=Low, show Graph<br>3. If SessionPctCorrect=High, show Illus<br>4. If ProblemPctCorrect=High, show Illus<br>5. If PctThruProblem=Early, show Graph<br>6. If PctThruSession=Early, show Graph | 1. If SessionPctCorrect=High, show Graph<br>2. If PctThruSession=Early, show Illus<br>3. If ProblemPctCorrect=High, show Illus<br>4. If PctThruProblem=Early, show Illus<br>5. If RQPctCorrect=Low, show Illus |
| **Male High Pretesters (n = 3)** | **Male Low Pretesters (n = 9)** |
| 1. If RQPctCorrect=Low, show Illus<br>2. If SessionPctCorrect=High, show Illus<br>3. If WalkThruPctCorrect=Low, show Illus | 1. If RQPctCorrect=Low, show Illus<br>2. If WalkThruPctCorrect=Low, show Illus<br>3. If SessionPctCorrect=High, show Illus<br>4. If PctThruSession=Early, show Graph<br>5. If PctThruProblem=Early, show Graph<br>6. If ProblemPctCorrect=High, show Illus |

perform significantly better than baseline, with PreScore*Gender performing significantly better than the rest; Table 2 has its policy.

In the model, we see differences between the partitions. Low pretesting females with a High PctSessionCorrect should be shown graphs, where as males and high pretesting females should be shown illustrations. When early in the tutoring session, low pretesting females should see illustrations whereas high pretesting females and low pretesting males should see graphs. When WalkThruPctCorrect is low, high pretesting females should see graphs whereas males should see illustrations. When RQPctCorrect is low, high pretesting females should see graphs but males and low pretesting females should see illustrations. That these differences exist in the model suggest that looking at interactions with both features improves situation description.

## 5 Discussion and Future Work

Prior work on this data found differences from the pooled data [11] by identifying when one group of students may benefit from one visual representation over another. This work identifies situations when one graphic might be better than the other for the same student and creates an adaptive model. In ongoing work, we have incorporated one model into a tutoring system and are evaluating its effectiveness at selecting visuals that aid learning compared to both alternating visual representations and using only one throughout tutoring.

This paper also presents a technique for mining data to create an adaptive policy when a gold standard is not available. It starts with a standard method (stepwise linear regression) and augments it to remove terms unwanted for developing adaptive systems. The method seeks to identify situations when one graphic is better than the other. Increasing situation descriptions, by adding more factors to each term, improve model performance. Many models, particularly those involving PreScore, significantly outperform the baseline. In ongoing

work, we are exploring improvements to model development, such as automatically identifying factors to add to a term to improve situational descriptions.

**Acknowledgments.** We thank Huy Viet Nguyen, Nathan Ong, and the Rimac team for their contributions. This research was supported by IES Grant R305A100163 to the University of Pittsburgh. The opinions expressed are those of the authors and do not represent the views of IES or the U.S. DoE.

## References

1. Aist, G., Kort, B., Reilly, R., Mostow, J., Picard, R.: Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: adding human-provided emotional scaffolding to an automated reading tutor that listens. In: IEEE International Conference on Multimodal Interfaces, pp. 483–490 (2002)
2. Arroyo, I., Beck, J.E., Park Woolf, B., Beal, C.R., Schultz, K.: Macroadapting animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. In: Gauthier, G., VanLehn, K., Frasson, C. (eds.) ITS 2000. LNCS, vol. 1839, pp. 574–583. Springer, Heidelberg (2000)
3. Chi, M., VanLehn, K., Litman, D.: Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. ITS (2010)
4. D'Mello, S.K., Graesser, A.: Affect detection from human-computer dialogue with an intelligent tutoring system. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 54–67. Springer, Heidelberg (2006)
5. Ekstrom, R., French, J., Harman, H., Dermen, D.: Manual for kit of factor-referenced cognitive tests. Educational Testing Service, Princeton (1976)
6. Forbes-Riley, K., Litman, D.: Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. Computer Speech & Language (2011)
7. Katz, S., Jordan, P., Litman, D., The Rimac Project Team: Rimac: A natural-language dialogue system that engages students in deep reasoning (2011)
8. Kohl, P.B., Finkelstein, N.D.: Student representational competence and self-assessment when solving physics problems. Phys. Rev. ST Phys. Educ. Res. (2005)
9. Kozhevnikov, M., Motes, M., Hegarty, M.: Spatial visualization in physics problem solving. Cognitive Science 31(4), 549–579 (2007)
10. Kulik, C., Kulik, J., Bangert-Drowns, R.: Effectiveness of mastery learning programs: A meta-analysis. Review of Educational Research 60(2), 265–299 (1990)
11. Lipschultz, M., Litman, D.: Illustrations or Graphs: Some Students Benefit From One Over the Other. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS (LNAI), vol. 7926, pp. 746–749. Springer, Heidelberg (2013)
12. McDermott, L., Rosenquist, M.: vanZee, E.: Student difficulties in connecting graphs and physics: Exampies from kinematics. American Journal of Physics (1987)
13. McNeil, N.M., Fyfe, E.R.: Concreteness fading promotes transfer of mathematical knowledge. Learning and Instruction, 440–448 (2012)
14. Meltzer, D.: Relation between students problem-solving performance and representational format. American Journal of Physics 73, 463 (2005)
15. Rau, M., Aleven, V., Rummel, N.: Intelligent tutoring systems with multiple representations and self-explanation prompts support learning of fractions. AIED (2009)
16. Van Heuvelen, A., Zou, X.: Multiple representations of work–energy processes. American Journal of Physics 69, 184–194 (2001)
17. VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Treacy, D., Wintersgill, M.: The andes physics tutoring system: Lessons learned. IJAIED (2005)