Learning Gains for Core Concepts in a Serious Game on Scientific Reasoning

Carol Forsyth, Philip Pavlik, Jr. Arthur C. Graesser, Zhiqiang Cai, Mae-lynn Germany
The University of Memphis
Institute for Intelligent Systems
400 Innovation Drive
Memphis,TN 38105
cmfrsyth,ppavlik,graesser,zcai mlgerman@memphis.edu

Keith Millis
Northern Illinois University
Psychology Building
Dekaulb,IL 60115
kmillis@niu.edu

Robert P. Dolan
Pearson Education
184 N Leverett Rd.
Leverett, MA 01054
bob.dolan@pearson.com

Heather Butler, Diane Halpern
Claremont Graduate University
Psychology Department
Claremont, CA 91711
hbutler,
diane.halpern@cmc.edu

ABSTRACT

OperationARIES! is an Intelligent Tutoring System that teaches scientific inquiry skills in a game-like atmosphere. Students complete three different training modules, each with natural language conversations, in order to acquire deep-level knowledge of 21 core concepts of research methodology (e.g., correlation does not mean causation). The student first acquires basic declarative knowledge and then applies the knowledge by critiquing case studies on scientific methodology and finally generating questions that reflect the core topics. A study using a pretest-training-posttest design was conducted in which 46 college students interacted with the modules of *OperationARIES!*, resulting in thousands of logged measures. The goal of this investigation was to discover the different trajectories of learning within 11 of these core concepts by evaluating 3 main constructs (e.g., discrimination, generation, and time on task) represented by key logged measures. Different constructs showed relationships with specific core concepts. Three core concepts were analyzed with stepwise regression and 5-fold cross-validation in order to discover contributing factors to learning gains for these core concepts.

Keywords

Intelligent Tutoring Systems, reasoning, serious games, research methods, discourse

1. INTRODUCTION

Social scientists often emphasize differences among students in their analyses of learning. The present research acknowledges such differences among students and aptitude-treatment interactions [1]. However the salient message in this study puts the magnifying glass on differences between core concepts in a subject matter. Simply put, the learning trajectories of core concepts may differ substantially depending on their content, complexity, and difficulty.

1.1 Cognitive Constructs Predicting Learning

The cognitive and learning sciences have identified principles of learning that offer likely hypotheses regarding differences in learning trajectories for core concepts [2]. Some concepts are learned by simply spending time reading and studying the material, a factor called time on task [3]. Time on task is normally optimized when concepts are presented on multiple occasions and distributed over time rather than concentrated in one time block [2, 4, 5]. Some concepts are learned primarily by actively generating the associated information about the concepts [2,4], particularly explanations [2, 5, 6, 7, 8]. Some concepts are best learned by testing experiences [9] and feedback on their answers [10], whereas others are best learned by either tutorial interaction [8, 11, 12, 13], scaffolding to get the student to generate good questions about difficult conceptualizations [14, 15], or tasks to get the student to make important discriminations among alternatives [8, 11, 13, 14]. The present study investigates the training events and experiences that contribute to the acquisition of critical core concepts. Our central point is simple. Core concepts have idiosyncratic characteristics that lend themselves to particular learning activities that optimize their acquisition.

The goal of this investigation is to discover the cognitive factors that predict the learning of core concepts in research methodology. The concepts range from concrete to abstract topics [8, 11, 14] and may require the student to utilize different skills. For example, understanding the meaning of an operational definition may be quite shallow in nature and possibly only require more time on task. Conversely, a more challenging abstract topic such as

correlation vs. causation may not be mastered by simply memorizing a definition but rather by higher level reasoning, discrimination among similar constructs, and generating ideas or questions. The learning environment is a serious game called Operation ARIES, as described next [12]. Although we have considered thousands of measures collected during 20 hours of training in ARIES, our analyses converged on three broad time-honored constructs in the cognitive and learning sciences: time on task, discrimination, and generation.

1.2 Operation ARIES: A Serious Game

OperationARIES! (called ARIES for short, an acronym meaning Acquiring Research Investigative and Evaluative Skills) is an Intelligent Tutoring System that has an embedded storyline and game-like elements to engage students as they learn research methodology. The narrative includes alien invaders who have come to take over the world by presenting bad science. The student player joins forces with the Federal Bureau of Science in order to save the world from this threat. The storyline and the iterative presentation of these topics are presented to the students across three specific ARIES modules (i.e., Training module, Case Study module, and Interrogation module), each focusing on different types of knowledge acquisition: didactic knowledge, application, and question generation. The learner interacts in natural language conversations with multiple artificial agents in order to learn 21 core concepts of research methodology.

In the Training module students learn didactic knowledge by reading an E-text, answering multiple choice questions, and having dynamic tutorial conversations with two pedagogical agents about the 21 core concepts. In the Case Study module, students apply the knowledge by conversing with three artificial agents while identifying flaws in research cases with the aid of both a list of 12 potential flaws and the E-book. Finally, in the Interrogation module, students pose questions to an artificial agent in order to decide if the research case is sound. The learner is aided by a score-card which provides immediate feedback as well as suggested questions. The flaws covered in the Case Study module and Interrogation module are aligned with the core concepts in the Training module.

This paper explores the specific cognitive activities in this serious game that predict learning of a subset of the 21 core concepts. These cognitive activities are part of the Training, Case Study, and Interrogation modules.

2. METHODS

The participants were 46 students at 2 separate schools in Southern California. There was a pretest-training -posttest design, with two versions of a test that were counterbalanced between pretest and posttest. All of the students were enrolled in research methodology courses taught by the same instructor. The pretest and posttest consisted of open-ended and multiple-choice questions about the 21 core concepts. The participants interacted with the Training module in pairs, alternating between actively typing into the system and passively observing their human partner interacting (a difference that was not analyzed in this study). The participants intermittently answered survey questions about the storyline and tutorial conversations, but these measures are not investigated in the current study. The alternation between partners as well as the surveys did not occur in the latter two modules (Case Study and Interrogation).

2.1 Measures

The log files of ARIES had thousands of measures including fine-grained measures for each module. Measures include latency measures, string variables and virtually every aspect of the typed interaction. With so many variables, the focus of this particular investigation will be on those measures that funnel into the three constructs of time on task, generation, and discrimination.

Each of the 3 constructs was represented by a unique indicator for each module. Specifically, time on task was represented in the Training module by reading times per page in the E-Text, whereas the time spent on cases was the measure for the Case Study and Interrogation modules. In order to assess generation, the measures consisted of the number of words articulated by the student in conversational turns for each module. Discrimination scores were collected in each module. The Training module used the multiple-choice performance scores (0 to 1). In the Case Study module, a discrimination score was calculated by subtracting the proportion of false alarms from hits as reflected by the match scores of the language processing algorithms within the system. The Interrogation module also used signal detection components derived from student performance on the score-card that discriminated whether a flaw was or was not present in a study.

In order to measure learning gains, we computed proportion scores for the pretest and posttest. Each test consisted of a multiple-choice and short-answer question corresponding to each of the 21 concepts. Proportional learning gains scores [(posttest-pretest)/(1-pretest)] were calculated in order to adjust for the variation of prior-knowledge across the students. These scores were available for each of the 21 concepts.

3. ANALYSES

Although this original dataset consisted of 46 participants, 10 of the subjects were removed due to extensive amounts of missing data (i.e. usually more than one module). Of the remaining 36 students, mean values were used to replace the missing data for discrimination scores. However, time on task and generation scores were simply left as 0's.The most complete set of original data, prior to mean replacements were available for 11 core concepts.

These core concepts were presented and tested across all three modules, so they were selected in the subsequent analyses. .

3.1 Correlations

The proportional learning gain scores ranged from .17 (Causal Claims) to .50 (Subject Bias), with a mean of .34 over the 11 core concepts. We computed correlations between these gain scores and the training process measures. We found a number of significant correlations, but the more important conclusion is that the profile of process to learning correlations differed greatly among core concepts.

It is beyond the scope of this report to present the full set of data. Instead, we will focus on a few core concepts that illustrate the differences. For example, the Training module reigned in the learning of one core concept (Objective Scoring of the Dependent Variable) when inspecting the correlations, which were significantly positive for the three measures: reading time, words generated, and discrimination. In contrast, the Interrogation module was most important for Subject Bias, where the corresponding three measures had significant correlations.

The differences in learning process profiles among core concepts underscores our central claim that core concepts vary considerably in learning trajectories.

3.2 Stepwise Regressions and Cross-Validation

We performed analyses on three core concepts that had distinctive profiles of correlations. These included Objective Scoring, Subject Bias, and Causal Claims. Each of these core concepts was analyzed separately using stepwise regressions with predictor variables that included those with the highest correlations (r >|.2|) with proportional learning gains. The resulting model was then cross- validated using a 5-fold procedure with 4 folds for training and 1 for test..

3.2.1 Objective Scoring of the Dependent Variable

This core concept showed correlations with the proportional learning gains for the reading times (time on task measure, r = .32, p<.05) and the multiple choice questions (discrimination score, r = .32, p<.05) in the Training module. In all 3 modules, the number of words generated significantly correlated with proportional learning gains (Training (r = .42, p<.05); Case Study(r =.28, p<.05); Interrogation (r =.28, p <.05). When these significant correlates were entered into a stepwise regression, the analysis removed the time allocated to multiple choice questions (time on task) and the words generated in the Training module, thereby converging on a model that includes words generated in the Interrogation module and the Case Study module and the reading times from the Training module (F (3, 33) = 4.91, R^2 = .31, p<.05). In the full model, the words generated in the Interrogation module had a marginally significant main effect (F (3, 33) = 3.61, p = .06); the words generated in the Case Study module did not have a significant main effect (F (3, 33) = 2.45, p = .13), but reading times were significant (F (3,33) = 8.67, p<.05). Given these results, a second model was created using the generation score for the Interrogation module and the reading times. The model was significant (F (2, 34) = 4.338, R^2= .20, p<.02) with a marginally significant main effect for generated words (F (2, 34) = 3.23, p = .08) and a significant main effect for reading times (F (2, 34) = 5.45, p<.05). When this model was cross validated, the training set accounted for 26% of the variance (R^2= .26), and a test set accounted for 25% of the variance (R^2 = .25)

3.2.2 Subject Bias

For this core concept, the variables with the highest correlations with learning gains were the multiple choice discrimination score from the Training module (r =.20, p<.10), and the discrimination (r = .20, p<.1), generation (r =.33, p<.05), and time on task (r = .26, p<.05) measures from the Interrogation module. With all predictors entered into a stepwise regression, the resulting significant model included only the words per case (generation) and the discrimination score from the Interrogation module (F (2, 34) = 3.304, R^2= .16, p<.05). Upon further examination, there is a significant main effect for generation (F (2, 34) = .498, p<.05) but not for the discrimination score (F =1.63, p>.05). A second linear model with just the generation score was significant model (F (1, 35) = 4.368, R^2= .11, p <.05). Next, the significant generation predictor only was cross-validated using a 5-fold cross validation procedure resulting in a training set predicting 8% of the variance (R^2 =.08) and a test set predicting 6% of the variance (R^2 = .06). However, we are still tentative about drawing strong conclusions from this because of the low power in detecting differences in the regression.

3.2.3 Causal Claims

This core concept had low learning gains (.17) compared with the other topics. The two variables with highest correlations for learning were discrimination from the Case Study module (r = .28, p<.05) and the generation metric in the Interrogation module (r = .23, p <.1). However, a follow-up analysis with stepwise multiple regression was only marginally significant (F (2, 34) = 2.863, R^2 = .14, p=.07) and cross validation assessments were not significant.

4. CONCLUSIONS

Our analyses revealed very different learning profiles for specific core concepts in research methodology. The value of the didactic Training module was most pronounced for Objective Scoring of Dependent Variables, whereas the Interrogation module was most successful for Subject

Bias, and Case Study was most promising for Causal Claims. The constructs of time on task, generation of information, and discrimination were also quite different for the different core concepts. Moreover, students did not learn much about differentiating causal from correlational claims. This topic may be very abstract to many students, difficult to comprehend, and in need of substantially more training.

One important implication of this study is that the different core concepts might be assigned different modules or a different amount of training allocated to each module. For some core concepts, it may be sufficient to have them read text and prompt them to articulate propositions in language. For other core concepts, they need a large number of case study examples to apply their knowledge in a discriminating fashion. Simply put, training experiences need to be optimally allocated to the constraints of content.

There are a number of limitations in this study that prevent us from making more definitive claims about the type of training that should be matched to our core concepts. The study had a low number of participants and a moderate number of missing values for observations. However, we can confidently state that correlations between learning gains and the key constructs of generation, discrimination and time on task do vary across core concepts of research methodology in *OperationARIES!*.It is important to explore different learning trajectories of specific core concepts in addition to differences among students.

5. ACKNOWLEDGEMENTS

This research was supported by the Institute for Education Sciences (R305B070349) and National Science Foundation (HCC 0834847). The opinions expressed are those of the authors and do not represent views of the IES and NSF.

6. REFERENCES

[1] Cronbach, L. and Snow, R. Aptitudes and Instructional Methods: A Handbook for Research on Interactions. New York: Irvington, 1977.

[2] Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., and McDaniel, M. Organizing Instruction and Study to Improve Student Learning: IES Practice guide. (NCER 2007-2004). Washington, DC: National Center for Education Research, 2007.

[3] Taraban, R., Rynearson, K., and Stalcup, K. Time as a variable in learning on the world wide web. Behavior Research Methods, Instruments, & Computers, 33 (March 2001), 217-225.

[4] Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T. and Rohrer, D. Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological Bulletin, 132 (September 2006), 354-380.

[5] Kopp, K., Britt, A., Millis, K., and Graesser, A. Improving the efficiency of dialogue in tutoring. Learning and Instruction, (in press).

[6] McNamara, D. S., O'Reilly, T. P., Best, R. M., and Ozuru, Y. Improving adolescent students' reading comprehension with iSTART. Journal of Educational Computing Research, 34 (June 2006), 147-171.

[7] Roscoe, R.D. and Chi, M.T.H. Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. Review of Educational Research, 77 (December 2007), 534-574.

[8] VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., and Rose, C.P. When are tutorial dialogues more effective than reading? Cognitive Science, 31 (January 2007), 3-62.

[9] Roediger, H. L. III. and Karpicke, J. D. The power of testing memory: Basic research and implications for educational practice. Psychological. Science, 1 (September 2006), 181-210.

[10] Shute, V. J. Focus on formative feedback. Review of Educational Research, 78 (March 2008), 153-189.

[11] Graesser, A.C., Conley, M., and Olney, A. Intelligent tutoring systems. In K.R. Harris, S. Graham, and T. Urdan (eds.), APA Educational Psychology Handbook: Vol. 3. Applications to Learning and Teaching(2012). Washington, DC: American Psychological Association, 451-473.

[12] Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., Halpern, D. Operation ARIES!: A serious game for teaching scientific inquiry. In M.Ma, A. Oikonomou & J. Lakhmi (eds.) Serious Games and Edutainment Applications (in press), Springer-Verlag, London, UK.

[13] Ritter, S., Anderson, J. R., Koedinger, K. R., and Corbett, A. Cognitive tutor: Applied research in mathematics education. Psychological Bulletin and Review, 14 (April 2007), 249-255.

[14] Graesser, A. C., Ozuru, Y., and Sullins, J. What is a good question? In M. G. McKeown and L. Kucan (eds.) Threads of coherence in research on the development of reading ability (2009), New York: Guilford, 112-141.

[15] Graesser, A. C., and Person, N. K.. Question asking during tutoring. American Educational Research Journal, 31 (Spring 1994),104-137.

About the authors:

Carol Forsyth is currently a PhD student in Experimental Psychology at the University of Memphis. She is also pursuing Cognitive Science graduate certification. Her main areas of interest include discourse processes, serious games, educational data mining, and Intelligent Tutoring Systems. Forsyth works as a graduate research assistant to Dr. Arthur Graesser, a Full Professor of Psychology at the University of Memphis, Co-Director of The Institute for Intelligent Systems, Senior Research Fellow at The University of Oxford and author of over 500 journal articles. In addition to Dr.Graesser, Forsyth has been fortunate to work with and learn from all of the co-authors on this paper.