Discovering Theoretically Grounded Predictors of Deep vs. Shallow Level Learning

Carol Forsyth,Arthur Graesser,
Philip Pavlik,Jr.
The University of Memphis
The Institute for Intelligent Systems
365 Innovation Dr.
Memphis, TN 38152
1 001 901 678 2000
cmfrsyth,graesser,
ppavlik@memphis.edu

Keith Millis
Northern Illinois University
Psychology Building
Dekaulb, IL 60115
1 001 815 753 7087
kmillis@niu.edu

Borhan Samei
The University of Memphis
The Institute for Intelligent Systems
365 Innovation Dr.
Memphis, TN 38152
1 001 678 2000
bsamei@memphis.edu

ABSTRACT

We investigated predictors of shallow and deep learning for 192 college students with high vs. low prior knowledge in a game-like intelligent tutoring system, OperationARA that has an eText, multiple-choice tests, case-based reasoning, and adaptive tutorial conversations. Students are expected to learn about 11 topics of research methodology across three modules that target factual information, application of reasoning to specific cases, and question generation. Our approach blends evidence-centered design (ECD) and educational data mining (EDM) methods to discover the best predictors of deep and shallow level learning for students of varying aptitudes within this game. Theoretically-grounded constructs (e.g., time-on-task, generation, discrimination) were found to be significant predictors of deep vs. shallow knowledge acquisition.

Keywords

Intelligent Tutoring Systems, evidence-centered design, learning, reasoning

1. INTRODUCTION

One major goal of computer-based learning sciences is to predict learning from behaviors and events in technology-based environments. Accomplishing this goal requires a mix of two schools of thought. First, *evidence-centered design* (ECD; [10]) proposes an accurate linking between theoretically-grounded constructs and observable measures. Second, *educational data mining* (EDM; [2]) suggests appropriate statistical modeling to discover phenomenon occurring in educational settings. The current investigation attempts to link these two important schools of thought while investigating learning within an Intelligent Tutoring System (ITS). Specifically, in line with evidence-centered design, well-researched cognitive constructs are investigated as predictors of learning at a fine-grained level. Educational data mining techniques make it possible to discover unexpected patterns on large scale data that may have nested factors, such as different students, instructors and classrooms.

The current study uses these techniques to investigate theoretically-grounded constructs (e.g. time-on-task, generation, and discrimination) as predictors of deep vs. shallow level learning for students with high vs. low prior-knowledge levels in an ITS known as Operation ARA.

1.1 Operation ARA

Millis and colleagues [9] created Operation ARA (previously known as *OperationARIES!*) with the hopes of increasing students' knowledge of research methodology in an environment that is adaptive to students' prior-knowledge levels and that is dynamic and engaging. Both game features and pedagogical techniques are incorporated in Operation ARA. However, the focus of the current study is on the pedagogical features occurring across three distinct modules: teaching students the basic factual information (Cadet Training), application of knowledge (Proving Ground), and question generation (Active Duty).

Across these three modules, students engage in different learning activities while learning 11 topics of research methodology (e.g. causation vs. correlation, random assignment). In the Cadet Training Module, students learn the basic didactic information about the topics via an E-text, multiple-choice questions and natural language tutorial conversations between the human student and two artificial agents. In the Proving Ground module, the student must apply the information learned in the Cadet Training module by identifying flaws in research cases with the help of agents and a hint list that includes a list of potential flaws. An example flaw is "the dependent variable is not valid", or "correlation was confused with causation". Finally, in the Active Duty module, learners actively generate questions about an abstract of a research case and judge the validity of the answer.

Thousands of measures are collected across the learning activities embedded within the three modules. This investigation incorporates pedagogical principles in the learning sciences to choose the measures that may have the most meaningful relationships with shallow and deep learning for college students that vary in prior-knowledge about research methods.

1.2 Well-Researched Cognitive Constructs

Cognitive psychologists have identified several performance metrics as well as cognitive and discourse constructs that predict learning for students in complex learning environments [11,14,15]. Shallow learning includes comprehension of explicit information whereas deep learning requires a mental model about the topics that can be applied to reasoning about cases. Separate constructs may correlate with deep vs. shallow learning at varying depths of processing. Evidence-centered design assumes that each of these hypothetical constructs is carefully aligned with the measures, events, and behaviors that are collected throughout the learning experience. The approach is to identify a small number of general constructs with theoretical underpinnings that are good candidates for predicting learning at the different depths of conceptual processing. The three constructs explored here are time-on-task, generation, and discrimination. These constructs are expected to have different weightings across topics, items, and students, which can be discovered with data mining techniques.

The time a student spends on any particular academic activity is referred to as *time-on-task*. Multiple empirical investigations substantiate a positive relationship between time-on-task and learning [4,14]. Varying degrees of time may be needed depending on the learner's prior-knowledge corresponding to the novelty of the information and the depth of processing on a shallow to deep-level continuum [5,13]. Taraban and colleagues [14] substantiate the positive relationship between time and task and learning, but also suggest that sensitive measures are required to discover the fine-grained relationships between time-on-task and learning.

Generation can be defined as the amount of words produced by students in their self-explanations, questions, and ideas articulated during learning. Beneficial effects of generation over passively reading have been reported in empirical investigations to increase deep learning [3,15]. The similarity theory [7], as well as the semantic associative memory (SAM) model [12], explains the generation effect by postulating a network of semantic associations that get activated during learning, with concepts activating semantically similar concepts in the network. The active generation of information both facilitates and is facilitated by the conceptual similarity of material, but sometimes at the expense of discriminating important distinctions and contrasts. Moreover, generation of information increases with greater organization of material based on prior knowledge [8] and greater depth of processing [5].

Discrimination can be described as separating the signal from the noise, or identifying a correct answer when provided with multiple alternatives. Students can obtain a deep-level conceptualization of difficult concepts through tasks that require them to discriminate between multiple alternatives [1,15] that require subtle distinctions. The theoretical underpinnings of this construct have been captured in the SAM model [12] as well as other models in traditional verbal learning and memory paradigms that focus on the distinctiveness versus similarity of information [7]. Hunt and McDaniel [7] suggest that distinctive items are more likely to be remembered in tasks that rely on recognition (corresponding to shallow knowledge) rather than recall of information (corresponding to deep-level knowledge), whereas similarity enhances performance in tasks that emphasize recall over recognition. However, the conceptual organization of the content must be specified for accurate predictions of performance in these memory paradigms.

The goal of the current investigation is to discover measures within the rich environment of Operation ARA representing all 3 of these time-honored constructs that predict shallow vs. deep learning considering the student's level of prior-knowledge and the topics studied across the three modules of Operation ARA.

2. METHODOLOGY

Participants included 462 students enrolled across 12 sections of research methods courses with 11 different instructors of an undergraduate Psychology course at Northern Illinois University. Students were expected to complete the game as part of the course curriculum, but they were not required to sign informed consents in compliance with the Institutional Review Board. Unfortunately, 232 participants were dropped because either they did not complete the consent form or had missing pretests or posttests. The data was further screened and revealed 38 participants to have extremely fast response times on either the pretest or the posttests. These participants were also excluded. The final number of participants was N = 192 across 11 classrooms and 9 instructors.

The study used a pretest-intervention-posttest design in which all students interacted with all of the modules of Operation ARA. The college students first completed one version of the assessment as a pretest. Next, the students interacted with the three modules of Operation ARA (i.e. Cadet Training, Proving Ground, and Active Duty). After completing the interaction, students completed the posttest.

2.3 Measures

2.3.1 Theoretically Grounded Constructs

Measures were calculated on a by-topic basis for each of the cognitive constructs (i.e. time-on-task, discrimination, and generation) within each of the three modules (i.e. Cadet Training, Proving Ground, and Active Duty). For all measures of time on-task, the square root of the overall metric was computed in order to achieve a normal distribution and accommodate diminishing returns from a gamma distribution with a long positive tail in the distribution. In the Cadet Training module, the measure for time-on-task was the square root of the time spent reading the E-text within the chapter. The measures for time-on-task in the Proving Ground and Active Duty modules were the square root of the total time spent per case for each module, respectively.

Discrimination was calculated for each module on a by-topic basis based on signal detection theory which compares correct answers from distractor information. In the Cadet Training module, discrimination was measured by performance on the multiple-choice questions within each chapter. In the Proving Ground and Active Duty modules, discrimination was scored by computing the proportional number of hits (correctly identified flaws) minus the proportional number of false alarms (incorrectly identified flaws).

Generation was calculated as the number of words produced by the student. Generation in the Training module was the overall number of words articulated by the student within each tutorial conversation per chapter. In the Proving Ground and Active Duty modules, the construct was represented by the total number of words generated by the student while articulating flaws or generating questions.

2.3.2 Assessment of Learning on a Topic Level

There were two versions of the pre- and post-test assessments (version A and version B), which were counterbalanced across students. Both versions included a total of 22 multiple-choice questions. There were two questions assigned to each topic, including a definition and applied question. The definitional questions were used as a measure of shallow learning, whereas the applied questions were a measure of deep learning. The fact that there were only two test items per topic would not provide a very sensitive measure. Therefore, the topics were clustered to gain a more reliable picture of the relationship between learning gains across the 11 topics. The topics were clustered in a previous study [6] based on learnability [(Posttest- Pretest)/ (1-Pretest)]/2 by using Multi-dimensional Scaling (ASCAL algorithm) that segregated topics into two groups (True Experiment vs. Sampling). The "True Experiment" cluster includes topics such as *control groups* and *random assignment* whereas the "Sampling" cluster includes topics such as *representative samples* and *subject bias*. Two topics were excluded because they did not fit into either cluster. In the current study, by using these two clusters, the 11 observations per participants were reduced to 2 groups.

After establishing the topic clusters, a proportional learning gains formula [(Posttest- Pretest)/ (1-Pretest)] was used to calculate learning for shallow and deep items to account for prior knowledge. For shallow and deep-learning gains, the proportional learning gains were also calculated independently for each topic cluster resulting in 4 PLG scores for each participant. Extreme negative values (PLG < -1) were removed from the data on an item level, which reduced the total number of items from 768 to 743 across the 192 participants.

3. ANALYSES AND RESULTS

Before performing any analyses, the measures were transformed using the Winsorizing method to ensure no outliers would skew the data. This method ensures that all outliers beyond 3 standard deviations above or below the mean of the z-score of the given measure are transformed to reflect endpoint scores. Next, two median splits were performed. The first separated the students into two groups based on prior-knowledge (i.e. high vs. low) for shallow learning gains. The second separated students into high and low-prior-knowledge for deep level learning gains. Therefore, the one participant could potentially be in a high-prior knowledge group for shallow learning and in the low- prior knowledge group for deep-level learning. This means that the four groups did not have an equal number of subjects as one subject could be in multiple groups, but rather the goal was to seek group equivalence. The final groups included: (Group 1) low prior-knowledge and shallow learning (N = 141 with 188 units of analyses), (Group 2) high prior-knowledge and shallow learning (N = 141 and 188 units of analyses), (Group 3) low prior-knowledge and deep learning (N = 141, 187 units of analyses), (Group 4), high prior-knowledge and deep learning (N = 126, 176 units of analyses).

Separate analyses were conducted for each of the 4 groups in the following stages. First, Pearson correlations were computed between the cognitive constructs and the PLG. Although this violates the assumption of independence of observations in correlation, these correlations are simply used as a guide. Next, a series of linear mixed-effect regression models were used to test models that included the highly significant correlates (r >|.2|) and also that accounted for the nested factors of participant, classroom, and instructor. The full models included the significant correlates as fixed factors and the random factors of participant, classroom, instructor as well as test form to account for counter-balancing test forms. The best fit models were then validated using 50 iterations of 4-fold cross-validation on the linear mixed fixed-random effects modeling using the R package "lme4" version 1.1-6 that was just released in 2014. Several of the random factors (i.e. participant, instructor, classroom) were not included in the cross-validation because equal distributions were not maintained across the training and test folds with the current dataset. A generalization proportion is also reported for each model. This is the proportion of the training-fold explained variance that generalizes to the test fold. These analyses were performed for each of the four groups (i.e. low knowledge and shallow learning, high knowledge and shallow learning, low knowledge and deep learning, and high knowledge and deep learning).

3.1 Low Knowledge & Shallow Learning

All of the potential predictors (the cognitive constructs for each module) were correlated with the shallow proportional learning gains (PLG) for students with low prior-knowledge. The analyses revealed a significant correlation between the discrimination metric in the Active Duty Module (referred to as ADdisc) with the PLG (r (189) = .24, p <.001).

The full model of ADdisc (i.e. discrimination in the Active Duty module) as a fixed factor with the 4 random factors of participant, classroom, instructor, and test was significantly different from the null model including only the random factors (X^2 (1) = 10.81, p <.001). The ADdisc accounted for about 5.2% of the variance above the random effects (R^2 =.052). The relationship between discrimination in the Active Duty module and PLG was positive in nature (β =.24, p <.001). This means that greater discrimination identifying flaws in the Active Duty module correlates with higher shallow proportional learning gains. The 4-fold cross validation of the mixed model including ADDisc as a fixed factor and test form as a random factor revealed a training set accounting for 6.2% of the variance and a test set accounting for 5.4% of the variance (R^2=.062 and R^2=.054, respectively). The generalization proportion was .86.

3.2 High Knowledge & Shallow Learning

The correlational analyses revealed strong correlations between Topic group (i.e. True Experiment vs. Sampling) and the number of words generated in the Proving Ground Module (referred to as PGwords) each significantly correlated with the shallow-level proportional learning gains (r(188)=.23, r(188)=.21, respectively).

The linear mixed-effects model with Topic group (i.e. Experimental vs. Sampling) and PGWords (generation in the Proving ground module) as fixed factors with the 4 random factors of participant, topic, and test was significantly different from the null model (X^2(2) = 16.67, p <.001) with the overall model accounting for 9% of the variance. Specifically, Topic Group accounted for about 5% of the variance (R^2 = .047) and PGWords accounted for 4% of the variance (R^2 = .039). Both the Topic Group and the words generated in the Proving Ground module had a positive relationship with proportional learning gains (β = .2, p <.01, β = .19, p <.01, respectively). The 4-fold cross validation of the full model including Topic Group and PGWords as fixed factors and test form as a random factor revealed a training set accounting for 10.5% of the variance and a test set accounting for 7.9% of the variance (R^2= .105 and R^2= .079, respectively) with a generalization proportion of .75.

3.3 Low Knowledge & Deep Learning

The Pearson correlations revealed a strong correlation between the discrimination metric in the Proving Ground module (referred to as PGDisc) as well as the time spent in the Active Duty module (referred to as ADTime) with the PLG (r =-.23, r =.24, respectively).

The mixed-fixed random effects model with PGDisc and ADTime as fixed effects and the 4 random effects was significantly different from the null model($X^2(2) = 16.99$, $p <.001$). The overall model accounted for about 8.5% of the variance ($R^2 =.085$) above the null model that included only the random factors with PGDisc accounting for 5.4% of the variance and ADTime accounting for 3.2% of the variance (R^2=.054, R^2=.032, respectively). Specifically, discrimination in the Proving Ground module was negatively correlated with learning whereas the time spent in the Active Duty module was positively correlated with learning

($\beta = -.18$, $p <.05$; $\beta =.18$, $p <.05$, respectively). The 4-fold cross validation of the full model including PGDisc and ADTime as fixed factors and test form as a random factor revealed a training set accounting for 9.2% of the variance and a test set accounting for 7.4% of the variance (R^2= .092 and R^2= .074, respectively) with a generalization proportion of .81.

3.4 High Knowledge & Deep Learning

Pearson correlations were performed between each of the constructs of interest and the proportional learning gains for the applied or deep questions. Unfortunately, no strong significant correlates were discovered. Therefore, the rest of the analyses were not conducted as the researchers concluded that a predictive model for deep-level learning for high-prior knowledge students could not be discovered from these data.

4. CONCLUSIONS

The investigation revealed significant models for three of the four groups of high versus low prior-knowledge for shallow versus deep learning. Specifically, discrimination in the Active Duty module (i.e. the question generation module) was the most predictive measure of shallow learning for students with low prior-knowledge. Word generation in the Proving Ground Module and sampling-oriented topics were positively correlated with shallow learning gains for high prior-knowledge students. The predictive model for students with low prior-knowledge suggested a negative relationship between discrimination in the Proving Ground module and deep-level learning gains as well as a positive relationship between the time spent in the Active Duty module (where students generate questions) and deep-level learning. Each of the models makes sense within the theoretical frameworks of the cognitive constructs used as predictors although they were not predicted a priori but rather discovered through educational data mining methods. Unfortunately, no predictors were found for high prior-knowledge students and deep-level learning. Perhaps good students with high-prior knowledge will achieve deep learning gains regardless of the tutorial experience.

There are limitations in this study. There could be a greater number of observations per prior-knowledge and deep versus shallow groups. There is also the possibility of other measures being better predictors of deep versus shallow level learning. A current investigation is underway to test multiple measures per construct and thereby determine the best predictors of deep vs. shallow level learning. Although there were limitations to this study, the overall results support the approach of blending evidence-centered design and educational data mining to conduct fine-grained investigations of student interactions within an Intelligent Tutoring System. Both are needed to identify when a particular learning principle will be effective for a particular topic and type of student.

5. ACKNOWLEDGMENTS

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] Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. 1995. Cognitive tutors: lessons learned. *J. Learn. Sci. 4*, 167-207.

[2] Baker, R.S.J.D., and Yacef, K. 2009. The state of educational data mining in 2009: A review and future visions. *J. Ed. Dat. Min*. 1, 3-17.

[3] Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., and Hausmann, R. G. 2001. Learning from human tutoring. *Cog. Sci.* 25, 471–533.

[4] Chickering, A. W., and Gamson, Z. F. 1987. Seven principles for good practice in undergraduate education. *AAHE Bull.* 39, 3-7.

[5] Craik, F.I. M. and Lockhart, Robert S. 1972. Levels of processing: A framework for memory research. *J. Ver. Learn. Ver. Beha*, 11, 671–684.

[6] Forsyth, C.M., Graesser, A.C., Cai, Z., Pavlik, P., Millis, K., and Halpern, D. 2013. Learner profiles emerge from a serious game teaching scientific inquiry. Presented at the *Annual Meeting of the American Educational Research Association*. (San Francisco, CA, April, 2013).AERA '13.

[7] Hunt, R. R., and McDaniel, M. A. 1993. The enigma of organization and distinctiveness. *J. Mem. Lang.* 3 ,421-445.

[8] Mandler, G. 1968. Organized recall. Individual functions. *Psych. Sci.* 13, 235-236.

[9] Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., Halpern, D. 2011. Operation ARIES!: A serious game for teaching scientific inquiry. In M.Ma, A. Oikonomou & J. Lakhmi (Eds.) Serious Games and Edutainment Applications (pp.169-196). London, UK. Springer-Verlag, 2011.

[10] Mislevy, R.J, Steinberg, L.S., and Lucas, J.F. 2003. On the structure of educational assessments.*Measurement: Interdisc. Res. and Pers*. 1, 3-67.

[11] Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger,K.,and McDaniel, M. 2007.Organizing instruction and study to improve student learning: IES practice guide Washington, DC: National Center for Education Research, 2004-2007.

[12] Raaijmakers, J. G. W., and Shiffrin, R.M. 1981. Search of associative memory. *Psych. Rev.* 88, 93-134.

[13] Simon, H. A. 1990. Invariants of human behavior. *Ann. Rev. Psych*. 41, 1-19

[14] Taraban, R., Rynearson, K., and Stalcup, K. 2001. Time as a variable in learning on the World Wide Web. *Beh. Res. Met., Instr., Comp*. 33, 217-225.

[15] VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., and Rose, C.P. 2007. When are tutorial dialogues more effective than reading? *Cog. Sci*. 31, 3-62.