Using Tutors to Improve Educational Games

Matthew W. Easterday, Vincent Aleven, Richard Scheines, and Sharon M. Carver

Human-Computer Interaction Institute, Carnegie Mellon University

Abstract. Educational games and tutors provide conflicting approaches to the assistance dilemma, yet there is little work that directly compares them. This study tested the effects of game-based and tutor-based assistance on learning and interest. The laboratory experiment randomly assigned 105 university students to two versions of the educational game Policy World designed to teach the skills of policy argument. The game version provided minimal feedback and imposed penalties during training while the tutor version provided additional step-level, knowledge-based feedback and required immediate error correction. The study measured students' success during training, their interest in the game, and posttest performance. Tutor students were better able to analyze policy problems and reported higher level of competence which in turn affected interest. This suggests that we can improve the efficacy and interest in educational games by applying tutor-based approaches to assistance.

Keywords: intelligent tutoring systems, educational games, policy argument, debate.

Educational games promise to make learning fun. Games typically provide less explicit assistance and harsher penalties than intelligent tutors, and perhaps as a result, more interesting choices. Do the mechanics that make games fun also promote learning? Or is lowered assistance the price we pay for increasing interest?

A review of educational game research shows a lack of empirical evaluation, especially the controlled comparisons between games and other approaches that would allow us to answer this question [1]. The explosion of AIED/ITS work on games in the last several years has produced scores of papers but has not radically altered the situation [2-4]. The great majority of work includes either no empirical evaluation or no control, and many of the remaining controlled experimental studies compare features that are important but not intrinsic to games or tutors [5-7].

To determine whether games offer a superior approach, we need to test whether their essential features, like the possibility of losing, the hidden or uncertain state created by opponents or random events, and the lack of external rewards, interfere with learning. We also need to test commonly used features like fantasy contexts [8]. Likewise we need to know whether the essential features of tutors such as how they provide step-level assistance interfere with interest. Of the few recent experimental studies in AIED/ITS, some have shown games to be inferior to or no better than didactic, non-intelligent instruction [9-10]. Another found no difference in learning, but a benefit in engagement for the game [11].

Tutors are inherently defined by how they provide assistance [12]. Thus, the greatest potential conflict between games and tutors is their different approaches to

G. Biswas et al. (Eds.): AIED 2011, LNAI 6738, pp. 63–71, 2011.

assistance. AIED/ITS research must determine whether games offer a superior solution to the assistance dilemma [13], or whether we simply transplant traditional tutor-based assistance into games without harming interest.

The purpose of this study was to compare the effects of *assistance* (using either a tutor or game-based approach) on learning and interest using an educational game called Policy World that teaches the skills of policy argument [14-15]. In the game-based version, the student received only a baseline level of assistance typically used in games including: situational feedback such as the game characters' dialogue, minimal-error flagging via the scoreboard, and penalties for making errors, such as restarting a level. In the tutor-based version, the student received additional knowledge-based feedback on every step and was required to immediately correct errors. In other words, the tutor always gave hints while the game let students die (fail and restart). Learning variables included students' learning of the search, comprehension, evaluation, diagram construction, synthesis, and decision skills taught by Policy World. Interest was measured by the Intrinsic Motivation Inventory [16].

These variables allow us to pose several competing hypotheses:

1. **Game hypothesis:** game-based assistance will increase learning and interest.
2. **Tutor hypothesis:** tutor-based assistance will increase learning and interest.
3. **Assistance tradeoff:** game-based assistance will increase interest, while tutor-based assistance will increase learning.

Intuitively, we might expect tutors to be more effective at increasing learning, because the principles upon which they are based have been derived from decades of empirical work [17] and because of the empirically demonstrated benefits of immediate, knowledge-based feedback with immediate error-correction [18]. On the other hand, situational feedback and delayed intelligent novice-feedback, similar to that offered by games, can be just as effective or even more effective at promoting learning as immediate, knowledge-based feedback [19-20], although their effects on interest are unclear. Intuitively, the game might be more fun because it gives the player more autonomy and the satisfaction of winning. On the other hand, excessive floundering is not fun, and the additional assistance offered by the tutor might be welcomed by a struggling student. These competing intuitions and tradeoffs form the core of the assistance dilemma especially as applied to educational games [13, 21].

Method

Population and Setting

105 university students were recruited through an on-line participant database and campus flyers. Students were compensated $20 for completing the on-line study, an additional $5 for passing posttest 1, and an additional $5 for passing posttest 2.

Intervention

Policy World. Policy World is an educational game designed for teaching the skills of policy argument. Students play the role of an analyst who must provide policy recommendations on topics like the drinking age, video game violence, carbon emissions, and national health care. The current version has 6 levels: a pretest, 3 training levels, and two posttests. Most levels include three broad activities: searching for policy

information, analyzing that information, and debating policy recommendations with a computer opponent. During search, students use a fake Google interface to find newspaper reports containing causal claims. During analysis, students use causal diagramming tools to analyze causal claims. During debate, students make a policy recommendation, explain how the policy will affect a desired outcome, and provide evidence for their position by citing reports. The analysis tools were disabled on the pretest and posttest 2.

Baseline game assistance. The baseline assistance in Policy World consisted of minimal feedback and penalties. During analysis, red/gold scoreboard stars indicated whether the student passed an analysis stage. During debate, the judge character would comment on critical mistakes and give the student a *strike* (as in baseball). The dialogue provided a form of situational feedback while the stars and strike provided a form of minimal feedback. As in most games, this assistance provided neither explicit teaching feedback nor was it at the step level. The baseline penalty was lost progress. When the student made an error on a stage of analysis of a particular causal claim, they were sent back to the first analysis stage. When the student received too many debate strikes, they had to replay the whole level. Note that in this study, students were automatically promoted after testing levels and were given the option to be promoted past a training level after playing it twice.

Tutoring assistance. The tutor version of Policy World provided supplemental assistance *only* during training. This additional assistance included explicit step-level feedback and immediate error correction. During training of analysis and debate the tutor provided explicit error-specific and teaching feedback on each step. The tutor also required immediate error correction, thus overriding Policy World's penalties.

Design

The study used a two-group, between-subjects, randomized, controlled, experimental design that compared a *game* to a *tutor* version of Policy World.

Task, Training Feedback, and Measures

Each Policy World level consisted of two phases: *search and analysis* and *debate*. In the search and analysis phase, the student searched for evidence using a fake Google interface to find 3-7 newspaper-like reports, 3-5 paragraphs in length, based on articles from sources like the *New York Times* and *Frontline.* At any time during this phase, the student could select a report to analyze which required him to comprehend, evaluate, diagram, and synthesize the evidence about the causal claims in the report.

Comprehend. After selecting a report to analyze, the student attempted to highlight a causal claim in the text such as: *the Monitoring the Future survey shows that 21 minimum drinking age laws decrease underage consumption of alcohol*.

Evaluate. The student then used combo boxes to identify the evidence type (*experiment, observational study*, *case*, or *claim)* and strength of the causal claim. Strength was rated on a 10-point scale with the labels: *none*, *weakest*, *weak*, *decent*, *strong*, and *strongest.* The evaluation was considered correct if: (a) the evidence type was correctly specified, and (b) the strength rating roughly observed the following order taught during training: experiments > observational studies > cases > claims.

Diagram. The student next constructed a diagrammatic representation of the causal claim using boxes to represent variables and arrows to represent an *increasing, decreasing,* or *negligible* causal relationship between the two variables. The student also "linked" the causal claim in the report to the new diagram arrow which allowed him to reference that report during the debate by clicking on that arrow.

Synthesize. The student then *synthesized* his overall belief about the causal relationship between the two variables based on all the evidence linked to the arrows between those variables up to that point. The synthesis step required the student to specify which causal relationship between the two variables was best supported by the evidence, and his confidence in that relationship on a 100 point slider from *uncertain* to *certain.* During training, a synthesis attempt was considered valid if: (a) the student moved his belief in the direction of the evidence, assuming the student's description of the evidence was correct, and (b) the student's belief mirrored the overall evidence, assuming the student's description of the evidence was correct.

Analysis feedback. During training, analysis errors resulted in animated red stars. Game students received no explanation for the error and were forced to restart the analysis of the claim. Tutor students received explanations and got to try again.

After ending the analysis phase, students moved to debate phase.

Recommendation. In the first step of the debate, the judge asked the student to choose a policy recommendation from a list of policy options which included increasing or decreasing any of the possible variables or *doing nothing*. For example, the student might recommend that: *we should repeal the 21 age drinking limit.* If the student proposed a recommendation that defied common sense or any directives given at the start of the problem, for example: *decreasing people's genetic propensity to drink,* the judge overruled the recommendation and gave the student a strike.

Mechanism. If the student proposed any recommendation besides *doing nothing*, the judge then asked the student to provide a mechanism that explained how the recommendation affected the desired policy outcome. The student used a set of combo boxes representing variables and causal relations to construct a mechanism such as: *repealing the drinking limit will decrease binge drinking which will decrease drunk driving*. If the student constructed an incoherent mechanism, for example that did not include the policy outcome, the judge gave the student a strike.

Mechanism Attack. If the student recommended *doing nothing*, the opponent proposed an alternate recommendation and mechanism, such as: *repealing the drinking limit will decrease binge drinking which will decrease drunk driving.* The student then had to attack a causal relation in the opponent's mechanism with an alternate relation, like: *repealing the drinking limit will not decrease binge drinking.* If the student made an incoherent attack by agreeing with the opponent or attacking a claim not in the opponent's mechanism, the judge gave the student a strike.

Evidence. After explaining or attacking a mechanism, the judge asked the student to cite reports with causal claims supporting the student's causal claim. Ideally, the student consulted his diagram by checking and clicking the relevant arrow on the diagram and checking the reports linked to that arrow during analysis. If the student provided a mechanism, the opponent would attack up to three causal claims in that mechanism before the student won the debate. If the student attacked an opponent's

mechanism, he only had to provide evidence for one attack, which would invalidate the opponent's entire causal chain. If the student provided irrelevant evidence, or weaker evidence than the opponent, the student received a strike.

Debate feedback. During training, all students received strikes for the gross errors described earlier. After 5 strikes, game students had to restart the level. Tutor students were given tutoring both for gross errors and any debate move inconsistent with their analysis. For example, a plausible recommendation inconsistent with the student's diagram would not receive a strike, but would receive tutoring. Citing sufficient, but not *all,* relevant evidence also initiated tutoring. Tutor students were then given Socratic tutoring on how to use the diagram and asked to try again.

Table 1. Posttest Measures

Measure	Description
Comprehend	# of claims correctly identified
Evaluate	# of correct evaluations of type and strength
Diagram	# of diagram elements linked to valid claims
Synthesize	# of times the synthesized relation and confidence shifted toward new perceived evidence and consistent the perceived evidence
Recommend	# of unique winning recs / (# unique winning recs + # losing recs)
Mechanism	# of unique winning mechanisms / (# unique winning mechanisms + # losing mechanisms)
Attack	# of unique winning attacks / (# unique winning attacks + # losing attacks)
Evidence	# unique winning ev. attempts / (# unique wining ev. attempts + # losing ev. attempts)
Training success	Average (# correct / (# attempts of each step on all training problems))
IMI	Intrinsic Motivation Inventory with sub-scales measuring: competence, effort, pressure, choice, value and interest [16]

During testing, students were allowed to construct arbitrarily incorrect diagrams, so we used the proxy measures of diagram and synthesis correctness.

Procedure

Students first took a *pretest* on either: junk food advertising and childhood obesity (13-15 causal statements), health care (8-9 causal statements), and cap and trade (9-10 causal statements). During the pretest, the analysis tools for comprehension, evaluation, construction, and synthesis were not available. All students were allowed to search for as many or as few reports as they liked before continuing to the debate.

Students were then randomly assigned to the *game* or *tutor* condition. Each group completed 3 *training* problems on video game violence (3 causal statements), the drinking age (12 statements), and the meth epidemic, (8 statements). During training, game students received only baseline feedback and penalties. Game students who failed a training level debate had to replay it once before being promoted to the next level. Tutor students received additional step-level, knowledge-based feedback and immediately corrected errors so they always successfully completed these levels.

Finally students played *posttest 1* (with analysis tools) and *posttest 2* (without analysis tools) levels counterbalanced with the pretest. The posttests provided neither the tutor nor the baseline analysis assistance. Students did not replay the posttests.

Results

Analysis 1: Who Learns More? We first examined success on the posttest 1 analysis steps. Table 2 shows that tutor students surpassed game students on every pre-debate analysis step. This suggests that adding step-level, knowledge-based feedback and immediate error correction increases learning in a game environment.

There was no significant difference between the two groups on any of the debate tasks on either posttest. By analogy to Algebra: tutor students did better constructing the equation but were still far from solving it.

Table 2. Comparison of Game and Tutor Groups on Posttest 1 Analysis

	Game		Tutor					
Measure	Mean	SD	Mean	SD	t	p	ll	ul
Comprehended	0.870	1.738	3.784	3.25	-9.816	6.18E-21 ***	-4.68	-3.1
Evaluated	0.704	1.449	2.549	2.48	-11.72	8.04E-28 ***	-5.03	-3.6
Diagramed	0.833	1.678	3.353	3.09	-5.148	2.00E-06 ***	-3.49	-1.5
Synthesized	1.056	1.937	5.078	4.42	-5.978	9.43E-08 ***	-5.37	-2.7

Analysis 2: Path model. We used path analysis to examine the causal relationships between assistance, training, interest, analysis, and debate. To search over all path models consistent with our background theories and that fit the data, we used the GES algorithm [22] implemented in Tetrad 4 to search for equivalence classes of un-confounded causal models consistent with the correlations in Table 3 and prior knowledge about the relationships between variables. This included the knowledge that: assistance was determined before any other factor, training was completed next, intrinsic motivation was measured before posttest 1, the student created a posttest 1 diagram before debating, and recommendations were provided before evidence.

Table 3. Correlations for Assistance, Diagramming, Debate, and Motivation

			Intrinsic Motivation Inventory										
	Assist	Train	Interest	Comp	Effort	Press	Choice	Value	Diag	Rec	Ev	M	SD
Asst	1											0.49	0.50
Train	.69***	1										0.58	0.19
Int	.05	.21*	1									3.82	1.34
Com	.44***	.50***	.57***	1								3.22	1.34
Eff	.04	.07	.13	-.04	1							5.16	1.06
Pres	-.14	-.22*	-.25***	-.37***	.34***	1						4.26	1.28
Cho	-.12	-.07	.44***	.36***	-.08	-.19*	1					3.58	1.08
Val	.12	.18.	.81***	.57***	.16	-.19*	.34***	1				4.37	1.35
Diag	.46***	.50***	.18	.42***	-.01	-.16.	-.02	.25***	1			2.07	2.77
Rec	-.01	-.01	.07	.18.	-.12	-.07	.06	.10	.26**	1		0.24	0.35
Ev	-.02	.07	.20***	.25***	-.02	-.13	.10	.19	.37***	.56***	1	0.22	0.37

*p<.05 **p<.01 ***p<.001

Figure 1 shows the model discovered by Tetrad which we consider highly plausible and shows an excellent fit to the data. A chi-squared test of the deviance of the path model from the observed values showed that we cannot reject this model at a significance level of .05, $\chi 2$ (40, n = 105) = 40.31, p > .46. Larger p-values indicate better fit and values above .05 indicate that we cannot reject the model at a significance level of .05.

According to the path model, tutor students had a greater success rate during training (as in Analysis 1). Students with greater success during training were more likely to diagram on posttest 1. Students who diagrammed more were more likely to make winning recommendations and to provide winning evidence. Students who had more success in providing recommendations were more likely to succeed in providing winning evidence. Those who received more assistance and those who had greater training success were more likely to report feeling competent. Those reporting higher competence valued the activity more for learning about policy, which increased interest. Those who perceived more choice while playing the game felt more competent and were more interested in the game, however assistance did not affect choice. Interest was correlated with, but did not cause competence and task success.

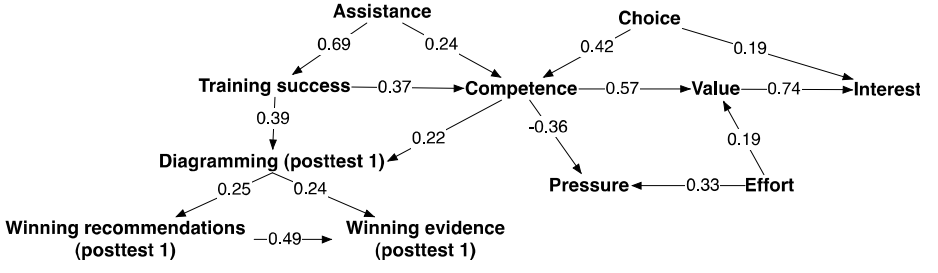

Fig. 1. A path model of the relations between the assistance, success on training, the amount of diagramming on posttest 1, posttest 1 debate performance, and intrinsic motivation

Discussion

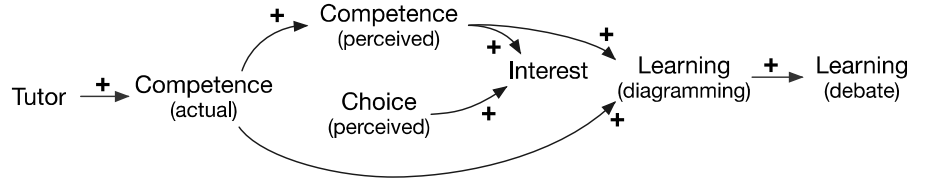

Fig. 2. Summary of results indicating support for mechanisms of the tutoring hypothesis

The results support the *tutoring hypothesis* that adding tutoring-based assistance to game environments increases both learning and interest (specifically competence). Figure 2 summarizes our view of the mechanisms that explain the patterns in the data we collected in this study. Adding tutoring to the game-like inquiry environment helped students succeed in training, which increased their ability to create diagrams on the posttest, which increased their ability to cite winning evidence during the

policy debate. Adding tutoring also increased students' self-reported competence, which increased their interest in the game which did not affect learning. Choice *did* increase interest in the activity, however choice was not affected by the tutor. The results can be described intuitively: assistance increased competence, which is good for learning and interest. The mechanisms between assistance, learning, and interest described by these results provide consistent support for the use of tutors in games.

Acknowledgements. This work was supported in part by a graduate training grant awarded to Carnegie Mellon University by the U.S. Department of Education (# R305B040063), the Siebel Scholars Foundation, and the Pittsburgh Science of Learning Center, funded by the National Science Foundation (# SBE-0836012). The opinions expressed are the authors' and do not represent the views of the U.S. Department of Education, the Siebel Scholars Foundation, or the National Science Foundation.

References

[1] Hays, R.T.: The Effectiveness of Instructional Games: A Literature Review and Discussion (Tech Rep. 2005-004). Storming Media (2005)
[2] Aleven, V., Kay, J., Mostow, J. (eds.): ITS 2010. LNCS, vol. 6094 & 6095. Springer, Heidelberg (2010)
[3] Dimitrova, V., Mizoguchi, R., du Boulay, B., Graesser, A. (eds.): Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling. IOS Press, Amsterdam (2009)
[4] Lane, H.C., Ogan, A., Shute, V. (eds.): Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education, Brighton, UK (2009)
[5] Ogan, A., Aleven, V., Kim, J., Jones, C.: Intercultural Negotiation with Virtual Humans: The Effect of Social Goals on Gameplay and Learning. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 174–183. Springer, Heidelberg (2010)
[6] Rowe, J.P., Mott, B.W., McQuiggan, S.W., Robison, J.L., Lee, S., Lester, J.C.: Crystal island: A narrative-centered learning environment for eigth grade microbiology. In: Lane, H.C., Ogan, A., Shute, V. (eds.) Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education, Brighton, UK, pp. 11–20 (2009)
[7] Lane, H.C., Hays, M.J., Auerbach, D., Core, M.G.: Investigating the Relationship between Presence and Learning in a Serious Game. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6094, pp. 274–284. Springer, Heidelberg (2010)
[8] Cordova, D.I., Lepper, M.R.: Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. Journal of Educational Psychology 88(4), 715–730 (1996)
[9] McQuiggan, S.W., Rowe, J.P., Lee, S., Lester, J.C.: Story-Based Learning: The Impact of Narrative on Learning Experiences and Outcomes. In: Woolf, B., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 530–539. Springer, Heidelberg (2008)
[10] Lane, H.C., Schneider, M., Albrechtsen, J.S., Meissner, C.A.: Virtual Humans with Secrets: Learning to Detect Verbal Cues to Deception. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010. LNCS, vol. 6095, pp. 144–154. Springer, Heidelberg (2010)

[11] Hallinen, N., Walker, E., Wylie, R., Ogan, A., Jones, C.: I was playing when I learned: A narrative game for French aspectual distinctions. In: Lane, H.C., Ogan, A., Shute, V. (eds.) Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education, Brighton, UK, pp. 117–120 (2009)
[12] VanLehn, K.: The behavior of tutoring systems. International Journal of Artificial Intelligence in Education 16(3), 227–265 (2006)
[13] Koedinger, K.R., Aleven, V.: Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review 19(3), 239–264 (2007)
[14] Easterday, M.W.: Policy world: A cognitive game for teaching deliberation. In: Pinkward, N., McLaren, B. (eds.) Educational Technologies for Teaching Argumentation Skills. Bentham Science Publishers, Oak Park (in press)
[15] Easterday, M.W., Aleven, V., Scheines, R., Carver, S.M.: Constructing causal diagrams to learn deliberation. International Journal of Artificial Intelligence in Education 19(4), 425–445 (2009)
[16] University of Rochester: Intrinsic motivation instrument, IMI (1994), http://www.psych.rochester.edu/SDT/measures/IMI_description.php
[17] Koedinger, K.R., Corbett, A.T.: Cognitive tutors: Technology bringing learning science to the classroom. In: Sawyer, K. (ed.) The Cambridge Handbook of the Learning Sciences, pp. 61–78. Cambridge University Press, Cambridge (2006)
[18] Corbett, A.T., Anderson, J.R.: Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In: Jacko, J., Sears, A., Beaudouin-Lafon, M., Jacob, R. (eds.) Proceedings of the ACM CHI 2001 Conference on Human Factors in Computing Systems, pp. 245–252. ACM Press, New York (2001)
[19] Nathan, M.J.: Knowledge and situational feedback in a learning environment for algebra story problem solving. Interactive Learning Environments 5(1), 135–159 (1998)
[20] Mathan, S.A., Koedinger, K.R.: Fostering the intelligent novice: Learning from errors with metacognitive tutoring. Educational Psychologist 40(4), 257–265 (2005)
[21] Aleven, V., Myers, E., Easterday, M., Ogan, A.: Toward a framework for the analysis and design of educational games. In: Biswas, G., Carr, D., Chee, Y.S., Hwang, W.Y. (eds.) Proceedings of the 3rd IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning, pp. 69–76. IEEE Computer Society, Los Alamitos (2010)
[22] Spirtes, P., Glymour, C., Scheines, R.: Causation, prediction, and search, 2nd edn. MIT Press, Cambridge (2000)