ABSTRACT

EXPLANATIVE AND ARGUMENTATIVE INTERACTIONS WITH AN INTELLIGENT TUTORING SYSTEM

by Colin Leigh Widmer

One-on-one tutoring has led to impressive learning gains in students. Tutors encourage students to elaborate answers, and elaboration has been shown to increase learning. Intelligent tutoring systems (ITS) are computer programs that simulate human tutors to interact with and encourage students to write elaborated answers. ITS and elaboration have primarily been used to teach well-structured domains with defined solutions. However, argumentation has been shown to increase learning in ill-structured domains where solutions are unclear. In the current study AutoTutor Lite, an ITS that uses natural language, was used to deliver lessons on the ill-structured domain of civics. The lessons included no interaction, an explanation interaction, argument interaction, or an irrelevant control. The lessons proved to be effective at teaching participants the material; however there was limited evidence that interacting provided additional benefits, though there is some indication that the kind of interaction with the tutor affected how participants value arguments.
EXPLANATIVE AND ARGUMENTATIVE INTERACTIONS WITH AN INTELLIGENT TUTORING SYSTEM

A Thesis

Submitted to the Faculty of Miami University
In partial fulfillment of
The requirements for the degree of
Master of Arts
Department of Psychology
by
Colin Leigh Widmer
Miami University
Oxford, Ohio
2013

Advisor ______________________________

Christopher Wolfe

Reader ______________________________

Cecelia Shore

Reader ______________________________

Joseph Johnson
# TABLE OF CONTENTS

INTRODUCTION .................................................................................................................. 1
  Tutoring and Intelligent Tutoring ................................................................. 1
  Argumentation and Ill-Structured Problems ............................................ 2

METHOD ................................................................................................................................. 3
  Participants .................................................................................................................. 3
  Materials .................................................................................................................. 3
  Procedure and Design .................................................................................. 4

RESULTS ............................................................................................................................ 5

DISCUSSION .......................................................................................................................... 8

REFERENCES .................................................................................................................... 11

APPENDIX A: Declarative Knowledge Test ...................................................... 17
APPENDIX B: Argument Rating Task ............................................................... 23
APPENDIX C: Essay Questions ............................................................................ 28
APPENDIX D: Essay Scoring Rubrics................................................................. 29
LIST OF TABLES

Table 1 ......................................................................................................................... 13
Table 2 ......................................................................................................................... 13
Table 3 ......................................................................................................................... 13
Table 4 ......................................................................................................................... 14
Table 5 ......................................................................................................................... 14
Table 6 ......................................................................................................................... 14
Table 7 ......................................................................................................................... 14
Table 8 ......................................................................................................................... 15
Table 9 ......................................................................................................................... 15
Table 10 ....................................................................................................................... 15
Table 11 ....................................................................................................................... 15
LIST OF FIGURES

Figure 1 ........................................................................................................................................ 16
INTRODUCTION

Tutoring and Intelligent Tutoring

The use of one-on-one tutoring has been shown to lead to very large learning gains. A well trained human tutor can enable a student to learn much more effectively than that same student would from studying alone. This increase in learning can be quite dramatic, with students that are tutored performing up to two sigma above students’ learning in conventional classrooms for highly trained tutors (Bloom, 1984; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2005), and a more modest effect size of about 0.79 on average (VanLehn, 2011). Tutors offer a variety of benefits to their students that lead to this increase in learning. Tutors have the ability to engage a student’s attention, ask the student additional questions, and give the student feedback based on progress (Graesser & McNamara, 2010). However, what has been shown to be perhaps the greatest benefit of having a tutor is that tutors often encourage their students to elaborate on their knowledge. More often than not a student will begin to answer a question briefly, perhaps using only a single simple sentence. Tutors can push students to expand on this initial answer, encouraging them to develop towards a more elaborate answer that reflects a more complete answer. This self-explanation that is prompted by a tutor has been shown to lead to strong learning gains (Roscoe & Chi, 2008). Elaboration not only leads to the student having a deeper understanding of their knowledge, but it also allows the student to recognize any discrepancies or misconceptions they may have (Graesser & McNamara 2010).

A current challenge in the field of technology is to create advanced learning systems on a computer that can achieve these same learning gains that are seen by eliciting self-explanation. Successful learning systems would feature the same processes that lead to effective learning in one-on-one tutoring. To this end intelligent tutoring systems have been developed. Intelligent tutoring systems are computer based agents that interact with students in ways that are meant to simulate the experience of a student interacting with a human tutor (Graesser et al., 2004). The goal of an intelligent tutoring system is to emulate the elements of a skilled human tutor such that the interaction between the student and the computer will achieve the same learning gains. One of the most promising methods for intelligent tutoring systems to elicit self-explanation from students is to communicate with them using a natural language such as English. Intelligent tutoring systems that interact using natural language do so by comparing the text entered by a student to a set of expectation text provided to the tutor using methods such as latent semantic analysis (LSA) (Graesser et al., 2000). LSA is a computational technique to mathematically measure the semantic similarity of sets of texts. It accomplishes this by creating a semantic space from a large corpus of text. The semantic space is a mathematical representation of the semantic relationships among words based on their co-occurrences in the corpus (Landauer & Dumais, 1997). This works off the assumption that words that are frequently found together have related semantic meanings. The latent aspect of LSA comes into play as not only are words that occur frequently with each other represented as having a high level of semantic similarity, but words that occur frequently with the same others words are also represented as highly similar (Hu, Wiemer-Hasting, Graesser, & McNamara, 2007).

In the context of an intelligent tutoring system LSA is used to compare sentences entered by students to a body of text that includes the key concepts of a correct answer (Graesser, Chipman, Haynes, & Olney, 2005). The tutor can then give appropriate feedback to the student to encourage elaboration based on this comparison (Kopp, Britt, Millis, & Graesser, 2012).
Examples of intelligent tutoring systems that can interact with students using natural language through LSA are AutoTutor (Graesser et al., 2004), and the web-based AutoTutor Lite (Hu et al., 2009) that will be used in this study. AutoTutor Lite is a smaller intelligent tutoring system based on AutoTutor. AutoTutor Lite does not have all of the features of AutoTutor, but has been optimized to run online in a web browser. AutoTutor Lite is among the first intelligent tutoring system to run online, and the first web-based intelligent tutoring system to use natural language.

One limitation of the current knowledge of the utility of intelligent tutoring systems is that most research using intelligent tutoring systems has been done on the teaching of well-structured problems, such as physics problems and computer programming. The current study intends to begin to broaden this knowledge to more ill-structured domains where the goals, constraints, and legal operations are not well defined or understood.

**Argumentation and Ill-Structured Problems**

Ill-structured problems are problems where the goals, solutions, or rules are not explicitly defined (Fernandes & Simon, 1999). This is in contrast to well-structured problems, which are better defined and constrained. Well-structured problems are those where there are definite solutions that can be clearly classified as correct or incorrect, there is a set of constrained rules that can be applied to solve the problem, all the necessary information to solve the problem is accessible, and it can be modeled using only basic computing processes (Simon, 1973). In general, problems that involve mathematics or logic are well-structured problems. However, many complex problems lack some or all of these features and are thus ill-structured. In ill-structured problems, the goal may be unclear, some variables may not be directly predictable or may interact in unpredictable ways, and solutions are difficult or impossible to classify as clearly correct or incorrect, with solutions often relying on an argument to justify that the solution is effective in solving the problem (Voss, 2005).

Because justification by an argument is a common feature of the solution to an ill-structured problem, (Voss, Tyler, & Yengo, 1983; Voss, 2005) it may be that learning via argumentation is appropriate in ill-structured domains. The creation of explanatory answers about material by a student has been shown to lead to an increase in learning about the material (Pressley et al., 1992), such as the elaboration that is encouraged by the interactions between a student and a tutor. This self-explanation has been shown to affect the learning of declarative knowledge from material (Chi, de Leeuw, Chiu, & LaVancher, 1994), but due to the nature of ill-structured domains the creation of an argument may be more effective in increasing learning that just the creation of an explanation when dealing with complex ill-structured material.

Argumentation is widely recognized as an important skill and is utilized in the education of a variety of fields (Wolfe, 2011), and the cultivation of argumentation skills in even young children is viewed as an important goal (Kuhn, 2010). There is also evidence that by creating arguments people achieve better understanding of material. Wiley & Voss (1999) found that those who were asked to write an argument when synthesizing historical material from multiple sources showed evidence of better understanding on several measures than those who wrote in the form of a summary, explanation, or narrative. Specifically, they showed greater ability on more complex tasks, such as recognizing inferences that followed from the material and recognizing the similarity of scenarios to the events studied, while on the more simple declarative recognition of material they performed more similar to those in other conditions.
Because of the results of this study it is reasonable to predict that using an intelligent tutor to teach students about an ill-structured domain and interact with them to create arguments about that domain may increase the depth of the students’ learning.

One domain that is wrought with ill-structured problems is civics. Most problems in the field of civics have multiple solutions that often undergo a great amount of argument. Competing ideas can operate side by side across States or countries, each with arguments supporting their use. For example, one ill-structured problem in civics is how judges should be selected. A commonly used solution is for judges to be appointed by an elected official, often the president or a governor. Proponents of this solution would have arguments supporting it, such as that because the offices held by judges are so important that it is essential that they be chosen by someone with expertise in the field of governing who has in depth knowledge of the requirements for office. However, another common solution is for judges to be elected by the general population. Proponents of this solution would have their own arguments to support it and counter the other solution. For example, they may argue that election by the people would keep elected judges accountable to the general population, while appointed judges may show loyalty to the individual or party that put them into office. Both solutions require arguments to support them, as neither is definitively more correct than the other. This and other problems in the domain of civics show that civics is an ill-structured domain that is useful for creating problems that can both be explained how they work, and be argued for a certain position.

There are three major hypotheses investigated in the present study. First, it is predicted that receiving a didactic lesson from the tutor will increase knowledge of a subject. Second, it is predicted that interacting with the tutor will benefit knowledge beyond receiving only a lesson. Interacting to create an explanation may benefit knowledge more than creating an argument, as writing an explanation may involve elaboration of more factual knowledge of the topic than an argument does. However, it is expected that either kind of interaction will increase knowledge over the lesson alone. Finally, it is predicted that interacting with the tutor to create an argument about an ill-structured domain will benefit more subtle knowledge of that domain than creating an explanation. It is expected that those who create an argument will have increased ability in tasks that require appreciating and using argumentation skills about that domain.

METHOD

Participants

123 Miami University undergraduate students participated in this study and received credit in an introductory psychology course in return.

Materials

AutoTutor Lite was used to deliver tutorial lessons in this study. Didactic lessons and tutorial interactions were developed in AutoTutor Lite using the available tutor development tools. The ill-structured domain that was chosen for the tutorial lessons was civics. Civics is a useful example of an ill-structured domain for creating tutorial lessons about problems that can be both explained and argued. Four lessons about civics were created for this study, each including a didactic lesson explaining the facts about the topic, a tutorial interaction based on explaining the topic, and a tutorial interaction based on creating an argument about the topic. The
four civics lessons created for the study were on the Electoral College, how State Supreme Court Justices are chosen, how State legislatures work, and term limits for governors. To emphasize the ill-structured nature of these topics, the lessons were focused on aspects of each topic that can be argued about (the Electoral College as opposed to the popular vote, whether judges are elected or appointed, the difference between a bicameral and unicameral State legislature, and governors with and without term limits).

Three tasks were developed to measure participants’ knowledge of the lessons and their ability to appreciate and use arguments about lessons. The first measure was a 40 item multiple choice tests on the declarative knowledge of the lessons which can be found in Appendix A. The second measure consisted of 48 arguments about the topics of the lessons for participants to rate how strong they though each argument was, as a measure of how they valued arguments of varying quality. The arguments in the measure were created to be of varying quality by including or excluding elements of the Toulmin Model of argumentation (Toulmin, 1958). The lowest quality arguments included only a claim (the conclusion of the argument). Intermediate quality arguments included the claim as well as a supporting reason (a fact that supports the claim). The highest quality arguments included a warrant (a statement justifying the support of the reason for the claim) in addition to the claim and reason. The arguments presented to participants can be found in Appendix B. The final task consisted of four short essays on the topics intended as a further assessment of participants’ knowledge and ability to use what they learned in the lessons. The essay questions can be found in Appendix C. These essays were scored for both knowledge shown by the participants and the participants’ use of argumentation using scoring rubrics that may be found in Appendix D.

Procedure and Design

Each participant received four lessons from AutoTutor Lite during the experiment, one in each of four conditions. The lesson each participant received with each condition was counterbalanced between subjects to create four replications of the study using the four lesson topics. In the first condition (lesson only) participants received a didactic lesson from AutoTutor Lite about one of the four civics topics. During this lesson participants were taught information by the tutor, but did not interact with the tutor in any way. In the second condition (explanation interaction) participants received another didactic lesson on one of the other civics topics. After this lesson was completed the tutor asked participants to create an explanative answer about what they had learned, interacting with the participants to give them feedback and improve the quality of their explanation of the material. In the third lesson participants received another lesson on a third civics topic. After this lesson participants interacted with the tutor to create an argumentation answer about the reasons for or against the sides of the topic they learned about. Again participants received feedback from the tutor to improve the quality of their argument. The content of the feedback in the explanation and argument interactions was kept constant, and the only differences between interaction conditions was whether feedback encouraged the writing of an explanation or an argument. A screenshot of the argument version of the tutorial interaction can be found in Figure 1. In the final condition (control), participants received a lesson about a topic unrelated to civics rather than a lesson of the fourth civics topic. In the control participants also did not interact with the tutor in any way during the lesson.

Each participant received each of the lessons in one of the four conditions (lesson only, explanation interaction, argument interaction, and control) with the lesson in the control
condition substituted with an unrelated lesson consisting of a portion of a lesson about genetic risk of breast cancer from a prior study using AutoTutor Lite (Wolfe et al., 2012). Participants were randomly assigned into four groups which counterbalanced the pairing of each condition with each lesson, as well as the order of the conditions (See Table 1 for a Latin Square of the design).

After completing the four tutorial lessons, all participants completed the three tasks. First participants answered the 40 item multiple choice test to assess their declarative knowledge of the content of the lessons. Next participants rated the 48 arguments about the topics on how strong they thought each was on a one through seven scale, as a measure of how they valued the arguments of varying quality. Lastly participants were asked to show and use what they had learned in answering the four short essay questions.

RESULTS

The results of each of the four lessons were analyzed in turn as four replications of the same study. Due to the design of the experiment each replication included the same participants, with each participant included in a different condition for each of the four replications. Using the same participants in each replication increases the likelihood that any differences found are primarily the result of the tutorial lesson manipulations. Any differences between the four problems themselves were not considered of theoretical interest.

As predicted, receiving a didactic lesson improved performance on the declarative knowledge tests on the factual knowledge of the topics. This result was found in all four lessons (Electoral College, F (3,119) = 15.541, p < 0.001; Judges, F (3,119) = 4.137, p = 0.008; State Legislatures, F (3,119) = 13.129, p < 0.001; Governors, F (3,119) = 5.382, p = 0.002). However, declarative knowledge appeared to be largely a function of the didactic lessons, as a planned Tukey HSD post hoc test revealed that all didactic lesson conditions (lesson only, explanation interaction, and argument interaction) performed significantly higher than the control condition at the 0.05 level. However, contrary to predictions there was no significant difference between any of the three groups that included a didactic lesson. This pattern can be seen in Table 2, which shows the means and standard deviations for all groups on the declarative knowledge test.

In order to assess participants rating of the strength of the arguments presented to them a difference score between their ratings of the stronger arguments (those that contained a claim supported by a reason and warrant) and their ratings of the weaker arguments (those that contained only a claim or only a claim and reason). Participants with a better appreciation of the strength of high quality arguments about a topic were predicted to show a greater sensitivity to the value of very well supported claims. Participants in the Electoral College lesson who interacted with the tutor to create an argument had a higher difference score between their ratings of high quality and low quality arguments than participants who interacted with the tutor to create an explanation, F (3,119) = 3.466, p = 0.018. That is, participants who wrote an argument valued arguments that supported their claims with both a reason and a warrant over arguments without such strong support more so than participants who had written an explanation, indicating that interacting to create an argument lead participants to become more sensitive to the role of warrants in supporting arguments. This follows prediction that interacting with the tutor to create an argument would enhance appreciation for more subtle knowledge about an ill-structured domain. However, this finding did not replicate in the other lessons (Judges, F (3,119) = 1.197, p
= 0.314; State Legislatures, F (3,119) = 0.822, p = 0.484; Governors, F (3,119) = 0.133, p = 0.940). A planned Tukey HSD post hoc test confirms that in the Electoral College the difference between the explanation interaction and argument interaction conditions was significant at the 0.05 level. Table 3 shows the means and standard deviations for all groups on the difference of rating the high and low quality arguments. Although not significant, it can be seen that in the Judges and State Legislature lesson do at least show a non-significant trend in the same pattern as seen in the Electoral College lesson, with those in the argument interaction condition appearing to show a greater difference than those in the explanation interaction.

The essays written by participants were scored using a rubric by two independent raters. The rubric included scoring the essays for knowledge of the topics covered in the lessons, whether the statements participants included in their essays about the topics were argumentative judgments, and whether the participants used their essays to argue a position on the topics. The independent raters had 87% agreement about the knowledge shown in the essays, 79% agreement on the inclusion of argumentative judgments in the essays, and 83% agreement on the positions taken by participants’ essays.

Again as predicted participants in all three groups that received a didactic lesson (lesson only, explanation interaction, and argument interaction) showed significantly more knowledge in their essays in the Judges lesson, F (3,119) = 2.710, p = 0.048, and the State Legislature lesson, F (3,119) = 5.272, p = 0.002. However, this difference was not found in the other two replications (Electoral College, F (3,119) = 0.674, p = 0.569; Governors, F (3,119) = 1.222, p = 0.305). This results appears to be also largely attributable to the didactic lessons, as again a planned Tukey HSD post hoc test shows that in both the Judges and State Legislature lessons the interaction conditions (both explanation and argument) showed greater knowledge than the control condition at the 0.05 level, but were not significantly different from each other or the lesson only condition. The lesson only condition was also significantly different from the control condition, but only in the State Legislature lesson. These differences can be seen in Table 4, which shows the means and standard deviations for all groups on the knowledge shown in essays. However, there was no evidence that interacting with the tutor has any benefit beyond the didactic lesson alone.

Essays were also analyzed based on the total number of statements participants included. A statement was defined as a unique and complete idea included in a participant’s answer. In the State Legislature lesson all participants who received a didactic lesson included a greater number of statements in their essay than participants in the control condition, F (3,119) = 6.228, p = 0.001, supporting the prediction that receiving a lesson would lead to better performance (in this case through increased writing). However, the effect was not replicated in the other lessons (Electoral College, F (3,119) = 1.397, p = 0.247; Judges, F (3,119) = 2.076, p = 0.107; Governors, F (3,119) = 1.418, p = 0.241). A planned Tukey HSD post hoc test showed that in the State Legislature lesson the difference between the three didactic lesson conditions and the control was significant at the 0.05 level but were not different from each other, indicating no evidence for benefits of interacting with the tutor above only receiving a lesson. This can be seen in Table 5, along with the means and standard deviations of all groups on the number of statements included in essays.

The proportion of the total number of statements included in the essay that made an argumentative judgment about the topic was also examined to determine how argumentative the
essays written by participants who receive different kinds of lessons were. Statements that make an argumentative judgment were defined as any statement that clearly supported or opposed any side of the topics, rather than just presented undisputed factual information about the topics from the lessons. There found to be no significant difference in the proportion of statements that were argumentative judgments in any lesson or group (Electoral College, F (3,119) = 1.289, p = 0.281; Judges, F (3,119) = 0.488, p = 0.692; State Legislatures F (3,119) = 1.652, p = 0.181; Governors, F (3,119) = 0.420, p = 0.739), indicating no support for the prediction that learning with the argument interaction would increase argumentation ability. Table 6 shows the means and standard deviations of the proportion of argumentative judgments in each group, revealing that there does not appear to be a strong consistent pattern between across lessons.

The essays were also classified based on whether the essay as a whole was taking an argumentative position on the topic as another measure of argumentation. A chi-square test found that participants who interacted to create an argument in the Electoral College lesson were significantly more likely to take an argumentative position than expected, $\chi^2 (1, N = 30) = 6.533$, $p = .006$. Additionally, participants who received the control lesson instead of the State Legislature lesson were less likely to take an argumentative position than expected, $\chi^2 (1, N = 30) = 8.533$, $p = .002$. However, there were no other differences in any other lesson or group, indicating limited support that learning with an argument interaction would increase argumentation ability. Table 7 shows the chi-square of the proportion of essays showing an argumentative position.

Unfortunately a number of participants did not complete the tutorial interactions as they were instructed. Some participants failed to complete the interactions, completing only a couple dialogue turns, instead of completing nine or ten if they had followed the tutor’s directions. Other participants skipped interactions entirely. Because interaction with the tutor is the key manipulation in this study it is problematic that participants failed to properly follow the tutor’s directions and could skew the results as these participants’ experience is similar or identical to those in the lesson only condition. In order to address this I conducted a post hoc analysis excluding participants who did not properly interact with the tutor. Participants were excluded if they did not interact with the tutor for at least four dialogue turns. Table 8 shows the number of participants included and excluded from each of the interaction conditions in the post hoc analysis.

The post hoc analysis replicated all of the significant findings of the original analysis and additionally revealed marginally significant trends in some of the replications that were not present when using the data of all participants. A planned contrast showed that in the Judges lesson there was a marginally significant difference in the rating of high quality and low quality arguments between the two interaction groups, with those who had interacted to create arguments valuing the additional support of a warrant more highly than those who had created explanations, just as was seen in the Electoral College lesson, $t (103) = 1.782$, $p = 0.078$, despite the overall effect not being significant, F (3,103) = 1.384, $p = 0.252$. Table 9 shows the means and standard deviations for the difference in rating the high and low quality arguments for all groups in the post hoc analysis. This slightly raises the amount of evidence that interacting with the tutor to create an argument increases sensitivity to the value of warrants in arguments. The pattern of the difference in ratings between the explanation interaction and argument interaction groups found in the original analysis is stronger in the post hoc analysis, remaining significant.
for the Electoral College lesson and approaching significance for the Judges pattern, although the Governors lesson still does not follow this pattern.

There was a marginal effect on knowledge shown in the essay in the Governor lesson, $F(3,105) = 2.474, p = 0.066$. This effect appears to be attributable to the explanation interaction, as participants in that conditions showed more knowledge than the others. This differs from the results for the Judges and State Legislature lessons from the original findings, where the effect could be attributed only to the didactic lesson. This difference can be seen in Table 10, which shows the means and standard deviations for all groups on the knowledge shown in essays in the post hoc analysis.

The post hoc analysis also revealed a marginal effect in the Judges lesson on the number of statements included in participant essays, $F(3, 103) = 2.59, p = 0.057$. Like in original findings for the State Legislature lesson participants in the explanation interaction and argument interaction groups included more statements than the control. However, they also include more statements that the lesson only condition, suggesting that the effect can be attributed to the interaction, rather than the didactic lesson alone. This can be seen in Table 11, which shows the means and standard deviations of the number of statements made in essays for the post hoc analysis.

**DISCUSSION**

There is strong evidence that all of the didactic lessons were effective at teaching participants the material. In all four replications the three groups that received the didactic lesson performed significantly better on the multiple choice declarative knowledge tests than the control condition. Because participants in the control condition did not show as much knowledge as those who did receive a lesson it is likely that most participants’ initial knowledge of the topics covered in the lesson was relatively low and that receiving the lessons actually improved knowledge. This evidence for the effectiveness of the lessons is particularly strong because the four replications included the same participants in different conditions, lowering the chances that any of the effect could be caused by variances in the level of knowledge participants in different conditions had of the domain of the lessons. In addition to this strong evidence, there is other sporadic evidence that the lessons themselves were effective, such as participants in the State Legislature lesson showing more knowledge and including more statements in their essays when they received a didactic lesson, though this did not carry over to the other replications.

Unexpectedly, unlike in previous studies in the literature (Graesser, McNamara, & VanLehn, 2005; Graesser et al., 2005), there was limited evidence in the present study that interacting with the tutor provided a benefit beyond that of the didactic lesson. Participants who interacted with the tutor showed no more knowledge on the multiple choice knowledge tests than those who only received the lesson without an interaction. There was only a small amount of evidence that supports a benefit of interaction with tutor, but only occurred in some of the replications, such as participants in the Governor lesson who interacted to create an explanations showing more knowledge in their essays and participants in the Judges lesson including more statements in their essays if they interacted with the tutor in either manner.

The specific interest of this study was to examine the differences between participants who interacted with the tutor to create an explanation and those who interacted to create an
argument. It was not expected that the kind of interaction would have much effect on participant knowledge, so it is unsurprising that there was no difference in performance on the declarative knowledge test. What was expected was that interacting to create an argument would affect performance on rating the strength of arguments and writing the essays. However, writing an argument appears to have had no effect on the essays, either in what information was included, how many argumentative judgments were made, or whether the participants used their essays to argue a position. Despite the absence of an effect on the essays, there does appear to be a potential effect on the argument rating task. Specifically, it appears that participants who previously interacted with the tutor to create and argument may have an increase in their sensitivity to the value of warrants in arguments when compared to participants who created an explanation. However, this effect is only clearly apparent in the Electoral College lesson. The effect does marginally replicate in the Judges lesson after participants who did not appropriately interact with the tutor were excluded, and this pattern of responses appears in the State Legislature lesson as well, though does not approach significance. One limitation of this finding is that there is a possible confound in the argument rating task. The arguments included were not controlled for length, and higher quality arguments tended to be longer. Thus the results could be alternately explained as an increase in sensitivity to the length of arguments. This confound can be removed in future research by controlling the length of arguments of all qualities.

There are a number of limitations of this study that may account for the failure to replicate the benefits of tutorial interaction and that there were only small differences between the explanation interaction and the argument interaction groups. One concern with the current study is participant motivation. This is most apparent with the answers to the essay questions. Despite being prompted to try to write a paragraph and include as much knowledge from the lessons as they could, on average participants wrote fewer than four statements and included only one or two out of many possible pieces of knowledge in their answers even in groups that received a lesson. Given the poor quality of the essay responses it is not surprising that few differences were found between groups. Additionally, although the tutor in total took participants about 45 minutes to complete, each individual lesson was only approximately 10 minutes. It is possible that the lessons were not long enough or did not include enough information for interaction to provide any additional benefit beyond what was learned during the lesson, at least at the level of motivation participants showed.

Another possible source of the failure to replicate the standard finding that interaction and self-explanation increases learning is the interactions themselves. It is possible that despite the careful development of the interactions, the tutor did not respond appropriately enough to participants’ responses. If the tutor was not able to give feedback to participants appropriately enough it is possible they did not fully engage in enough elaboration of their explanations and arguments for there to be enhanced learning beyond what was gained from the didactic lesson. The ability of AutoTutor Lite and other Intelligent Tutoring Systems to appropriately interact in natural language relies heavily on preprogrammed settings and feedback written in anticipation of participants’ responses. In previous research using AutoTutor Lite we have found a strong correlation between the appropriateness of the tutor’s responses to learners’ statements and performance on knowledge measures (Wolfe et al., 2013). The ability of subsequent versions of these tutorials to respond appropriately to participants can be improved based on the empirical data from this study. AutoTutor Lite keeps a record of each interaction between a learner and the tutor by creating a log that contains the exact script of what the learner entered and how the tutor
responded. This information can be used to both better predict what learners will enter and the kind of feedback that will help them. In addition, it provides a wealth of real answers entered by learners that can be used to test future iterations of the tutor to ensure it responds appropriately. Improvements in the appropriateness of the tutorial interactions may cause participants to engage in the interactions at a level that shows enhanced learning in future studies.

One direction to take to advance on this research would be to try to address some of these concerns by taking one lesson to develop further into a longer tutorial to use in further studies. The longer tutorial could contain more information about a topic that may reveal benefits to interaction with the tutor beyond what is gotten out of the lesson alone. Additionally, the interactions about that lesson can be more optimally developed to respond more appropriately to participants and better encourage self-explanation and argumentation. The increased amount of information in the lesson and optimally developed interactions may increase the chances of identifying any differences in the way participants learn the material based on creating explanations and arguments.

In addition to running a similar study using a more developed tutor there are other directions that could expand on this research. In the current study participants interacted with the tutor to create and argument but were provided with no training on argumentation or given an expectation about what makes a good argument. It is possible that including training on argumentation as an additional variable will have a stronger effect on how the material is learned. Participants in the current study were expected to write an argument about the topic of the lesson with the tutor providing feedback intended to help them include information to make a stronger argument. Another possible way to interact with an intelligent tutor to create an argument would be to argue with the tutor itself. If the tutor provided challenges and counterpoints to people’s argument and they then had to respond to these challenges it may cause people to process the information in a different way that might provide benefits beyond passively writing an argument with only informative feedback.

One other direction research on argumentation and intelligent tutoring could continue is to use intelligent tutoring systems to model argumentation and investigate vicarious learning through argumentation. Vicarious learning has been observed in intelligent tutoring by having a person observe a tutor–student dynamic between two talking avatars (Craig, Sullins, Witherspoon, & Gholson, 2006). The person watching the tutor is able to experience much of the benefit of tutoring without interacting directly through watching the student avatar ask questions and receive feedback from the tutor avatar. It is possible that similarly modeling argumentation between two avatars in an intelligent tutoring system would provide similar learning benefits.

In summary, there is strong evidence that the lessons effectively taught information to the participants as they were able to show greater declarative knowledge. However, interacting with the tutor appears to have had little to no effect on knowledge gain, despite the strong history of evidence that interaction with a tutor enhances learning. Future work with these tutorial lessons may uncover if the tutorial interactions themselves could be more effective, or if this discrepancy can be attributed to other factors like participant motivation or the amount of information in the lessons. Interacting with the tutor to create an argument rather than an explanation does not appear to have impacted knowledge, though there is some evidence that it may affect how participants judge and value arguments. Future research may reveal if this finding holds true.
REFERENCES


Table 1

*Experiment Design*

<table>
<thead>
<tr>
<th>Group</th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group I</td>
<td>Lesson Only</td>
<td>Explanation</td>
<td>Argument</td>
<td>Control</td>
</tr>
<tr>
<td>Group II</td>
<td>Control</td>
<td>Argument</td>
<td>Explanation</td>
<td>Lesson Only</td>
</tr>
<tr>
<td>Group III</td>
<td>Explanation</td>
<td>Lesson Only</td>
<td>Control</td>
<td>Argument</td>
</tr>
<tr>
<td>Group IV</td>
<td>Argument</td>
<td>Control</td>
<td>Lesson Only</td>
<td>Explanation</td>
</tr>
</tbody>
</table>

Table 2

*Declarative Knowledge Test Percent Correct (Means and Standard Deviations)*

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson Only</td>
<td>.650 (.198)</td>
<td>.480 (.179)</td>
<td>.613 (.194)</td>
<td>.518 (.205)</td>
</tr>
<tr>
<td>Explanation Interaction</td>
<td>.613 (.194)</td>
<td>.573 (.227)</td>
<td>.636 (223)</td>
<td>.550 (.216)</td>
</tr>
<tr>
<td>Argument Interaction</td>
<td>.623 (.201)</td>
<td>.561 (.215)</td>
<td>.653 (.185)</td>
<td>.493 (.168)</td>
</tr>
<tr>
<td>Control</td>
<td>.376 (.137) **</td>
<td>.417 (.168) **</td>
<td>.373 (.191) ***</td>
<td>.367 (.165) **</td>
</tr>
</tbody>
</table>

** The mean is significantly different from the other groups at the .01 level
*** The mean is significantly different from the other groups at the .001 level

Table 3

*Average Difference in Ratings of Arguments (Means and Standard Deviations)*

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson Only</td>
<td>1.1 (1.04)</td>
<td>0.34 (0.62)</td>
<td>0.50 (0.92)</td>
<td>0.53 (0.94)</td>
</tr>
<tr>
<td>Explanation Interaction</td>
<td>0.53 (1.15) *</td>
<td>0.43 (0.86)</td>
<td>0.33 (1.00)</td>
<td>0.42 (0.76)</td>
</tr>
<tr>
<td>Argument Interaction</td>
<td>1.47 (1.14) *</td>
<td>0.75 (0.94)</td>
<td>0.52 (0.98)</td>
<td>0.35 (1.12)</td>
</tr>
<tr>
<td>Control</td>
<td>0.98 (1.29)</td>
<td>0.53 (1.01)</td>
<td>0.26 (0.81)</td>
<td>0.39 (0.92)</td>
</tr>
</tbody>
</table>

* The means are significantly different from each other at the .05 level
Table 4

Knowledge Items Included in Essay (Means and Standard Deviations)

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson Only</td>
<td>1.80 (1.16)</td>
<td>1.23 (1.17)</td>
<td>2.60 (1.61)**</td>
<td>2.30 (1.53)</td>
</tr>
<tr>
<td>Explanation Interaction</td>
<td>1.70 (0.92)</td>
<td>1.73 (1.39)*</td>
<td>2.67 (2.13)**</td>
<td>2.90 (1.35)</td>
</tr>
<tr>
<td>Argument Interaction</td>
<td>1.93 (1.09)</td>
<td>1.88 (1.54)*</td>
<td>2.97 (1.19)**</td>
<td>2.37 (1.33)</td>
</tr>
<tr>
<td>Control</td>
<td>1.58 (1.00)</td>
<td>1.07 (1.11)</td>
<td>1.40 (1.48)</td>
<td>2.50 (1.11)</td>
</tr>
</tbody>
</table>

* The mean is significantly higher than the control condition at the .05 level
** The mean is significantly higher than the control condition at the .01 level

Table 5

Number of Statements in Essays (Means and Standard Deviations)

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson Only</td>
<td>3.90 (2.17)</td>
<td>1.83 (1.60)</td>
<td>3.93 (2.10)***</td>
<td>3.76 (2.49)</td>
</tr>
<tr>
<td>Explanation Interaction</td>
<td>3.30 (1.80)</td>
<td>2.73 (1.46)</td>
<td>3.21 (2.23)***</td>
<td>3.47 (2.16)</td>
</tr>
<tr>
<td>Argument Interaction</td>
<td>4.13 (2.00)</td>
<td>2.76 (2.19)</td>
<td>3.27 (1.57)***</td>
<td>2.73 (1.68)</td>
</tr>
<tr>
<td>Control</td>
<td>3.30 (1.96)</td>
<td>2.00 (2.05)</td>
<td>1.87 (1.57)</td>
<td>3.50 (1.72)</td>
</tr>
</tbody>
</table>

* The mean is significantly higher than the control condition at the .001 level

Table 6

Proportion of Argumentative Judgments in Essays (Means and Standard Deviations)

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson Only</td>
<td>.330 (.060)</td>
<td>.376 (.073)</td>
<td>.291 (.057)</td>
<td>.349 (.066)</td>
</tr>
<tr>
<td>Explanation Interaction</td>
<td>.352 (.064)</td>
<td>.423 (.049)</td>
<td>.289 (.045)</td>
<td>.294 (.061)</td>
</tr>
<tr>
<td>Argument Interaction</td>
<td>.455 (.055)</td>
<td>.378 (.063)</td>
<td>.414 (.057)</td>
<td>.343 (.059)</td>
</tr>
<tr>
<td>Control</td>
<td>.302 (.057)</td>
<td>.314 (.664)</td>
<td>.244 (.067)</td>
<td>.394 (.061)</td>
</tr>
</tbody>
</table>

Table 7

Proportion of Essays Taking an Argumentative Position (Chi Square)

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson Only</td>
<td>.533 (.133)</td>
<td>.500 (0)</td>
<td>.467 (.133)</td>
<td>.394 (1.485)</td>
</tr>
<tr>
<td>Explanation Interaction</td>
<td>.500 (0)</td>
<td>.567 (.533)</td>
<td>.515 (.030)</td>
<td>.467 (.133)</td>
</tr>
<tr>
<td>Argument Interaction</td>
<td>.733 (6.533)*</td>
<td>.515 (.030)</td>
<td>.567 (.533)</td>
<td>.600 (1.200)</td>
</tr>
<tr>
<td>Control</td>
<td>.424 (.758)</td>
<td>.400 (1.200)</td>
<td>.233 (8.533)*</td>
<td>.600 (1.200)</td>
</tr>
</tbody>
</table>

* The mean is significantly higher than the control condition at the .001 level
Table 8

*Number of Participants Excluded in Post Hoc Analysis*

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explanation</strong></td>
<td>26 (4 Excluded)</td>
<td>23 (7 Excluded)</td>
<td>21 (9 Excluded)</td>
<td>23 (7 Excluded)</td>
</tr>
<tr>
<td><strong>Argument</strong></td>
<td>26 (4 Excluded)</td>
<td>24 (9 Excluded)</td>
<td>21 (12 Excluded)</td>
<td>22 (8 Excluded)</td>
</tr>
</tbody>
</table>

Table 9

*Difference in Ratings of Arguments (Means and Standard Deviations)*

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lesson Only</strong></td>
<td>1.1 (1.04)</td>
<td>0.35 (0.62)</td>
<td>0.50 (0.92)</td>
<td>0.53 (0.94)</td>
</tr>
<tr>
<td><strong>Explanation Interaction</strong></td>
<td>0.64 (1.17) *</td>
<td>0.22 (0.81) †</td>
<td>0.29 (0.82)</td>
<td>0.52 (0.71)</td>
</tr>
<tr>
<td><strong>Argument Interaction</strong></td>
<td>1.42 (1.01) *</td>
<td>0.72 (0.94) †</td>
<td>0.49 (1.05)</td>
<td>0.50 (0.78)</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>0.98 (1.29)</td>
<td>0.53 (1.01)</td>
<td>0.26 (0.81)</td>
<td>0.35 (1.12)</td>
</tr>
</tbody>
</table>

* The means are significantly different at the .05 level
† The means are marginally different (p = .078)

Table 10

*Essay Knowledge Score (Means and Standard Deviations)*

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lesson Only</strong></td>
<td>1.80 (1.16)</td>
<td>1.23 (1.17)</td>
<td>2.6 (1.61) ***</td>
<td>2.30 (1.53)</td>
</tr>
<tr>
<td><strong>Explanation Interaction</strong></td>
<td>1.77 (.95)</td>
<td>1.74 (1.42)</td>
<td>2.71 (2.13) ***</td>
<td>3.22 (1.17) †</td>
</tr>
<tr>
<td><strong>Argument Interaction</strong></td>
<td>2.08 (1.02)</td>
<td>2.17 (1.58)</td>
<td>3.14 (1.15) ***</td>
<td>2.35 (1.47)</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>1.58 (1.00)</td>
<td>1.07 (1.11)</td>
<td>1.40 (1.48)</td>
<td>2.50 (1.11)</td>
</tr>
</tbody>
</table>

*** The means are significantly higher than the control condition at the .001 level
† The mean is marginally higher than the other conditions (p = .066)

Table 11

*Number of Statements in Essays (Means and Standard Deviations)*

<table>
<thead>
<tr>
<th></th>
<th>Electoral College</th>
<th>Judges</th>
<th>State Legislature</th>
<th>Governors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lesson Only</strong></td>
<td>3.90 (2.17)</td>
<td>1.83 (1.60)</td>
<td>3.93 (2.10) ***</td>
<td>3.76 (2.49)</td>
</tr>
<tr>
<td><strong>Explanation Interaction</strong></td>
<td>3.46 (1.86)</td>
<td>2.74 (1.36)</td>
<td>3.52 (2.27) ***</td>
<td>3.91 (2.15)</td>
</tr>
<tr>
<td><strong>Argument Interaction</strong></td>
<td>4.42 (1.84)</td>
<td>3.04 (2.26) †</td>
<td>3.33 (1.49) ***</td>
<td>2.65 (1.72)</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>3.30 (1.96)</td>
<td>2.00 (2.05)</td>
<td>1.87 (1.57)</td>
<td>3.50 (1.72)</td>
</tr>
</tbody>
</table>

*** The means are significantly higher than the control condition at the .001 level
† The means are marginally higher than the other conditions (p = .057)
Figure 1. Argument tutorial interaction. In the tutorial interactions the avatar provides feedback to participants as they create answers (in this case an argument about the Electoral College).
APPENDIX A: Declarative Knowledge Test

If two candidates receive the same number of votes in the Electoral College, the President is decided by…

A. The House of Representatives  
B. The Senate  
C. The Supreme Court  
D. Holding another election

A winter-take-all system is used by…

A. All States  
B. Most States  
C. Only Two States  
D. No States

Compared proportionally on population, low population States are...

A. Overrepresented in the Electoral College compared to high population States  
B. Underrepresented in the Electoral College compared to high population States  
C. Equally represented in the Electoral College compared to high population States  
D. State population does not affect electoral representation

In order to receive enough electoral votes to become president, a candidate needs…

A. To win the popular vote  
B. The electoral vote of at least half of the States  
C. 270 electoral votes  
D. All of the above

Faithless electors are held accountable by…

A. Fines in some States  
B. Fines in all States  
C. Jail time in some States  
D. Faithless electors are not held accountable in any State

The number of electors in each State is determined by…

A. The number of Senators plus Representatives the State has in Congress  
B. The State’s proportion of the national population  
C. The State’s constitution  
D. The number of votes cast in each State
The minimum and maximum number of electoral votes a State may have is…

A. A minimum of 3 and a maximum of 55
B. A minimum of 3 and no maximum
C. No minimum and a maximum of 55
D. No minimum or maximum

Which of the following non-States have electoral votes?

A. Washington D. C.
B. Puerto Rico
C. U.S. Virgin Islands
D. All of the above

The official electoral vote is held…

A. At the end of Election Day
B. About a month after Election Day
C. Just before the inauguration
D. After the new year

A consequence of the Electoral College is that most elections are decided by…

A. Mostly large States
B. Mostly small States
C. Mostly solid States
D. Mostly swing States

Court decisions by State Supreme Courts are appealed to…

A. The Supreme Court of the United States
B. United States Courts of Appeals
C. United States District Courts
D. State Supreme Court decisions cannot be appealed

In States that hold elections for judges…

A. Judges are elected in nonpartisan elections
B. Judges are elected in partisan elections
C. Judges are elected in only special elections
D. It varies by State

Governors have absolute authority to appoint judges without the approval of legislature or committees in

A. No States
B. One State
C. Three States
D. Eight States
Which of the following is not a method used to select judges in any State?

A. Nomination by a commission and appointment by governor
B. Direct election by the people
C. Election by the State Legislature
D. Election by the State Bar Association

The commission used in the traditional Missouri Plan does not include which of the following as members?

A. Citizens appointed by the Governor
B. Lawyers elected by the State Bar Association
C. The current State Chief Justice
D. Members of the State Senate

The Missouri Plan was developed with the goal of…

A. Appointing the most qualified candidates
B. Appointing the most long serving candidates
C. Appointing political allies
D. Appointing the most popular candidates

Under the Missouri Plan, if the Governor does not choose a candidate in a timely manner…

A. The commission must nominate additional candidates
B. The commission may appoint a candidate
C. The legislature may appoint a candidate
D. The candidates are put to general election

Once a candidate is appointed under the Missouri Plan…

A. The candidate serves out the term
B. The candidate is subject to a retention election ever year
C. The candidate is subject to a retention election after one year of service
D. The candidate is subject to approval by the State Bar Association

In the hybrid form of the Missouri Plan, candidates selected by the Governor are…

A. Subject to approval by the legislature
B. Subject to approval by the State Bar Association
C. Subject to approval in the general election
D. Subject to approval by nonpartisan commission

In bicameral State Legislatures, the upper house is called…

A. The Senate
B. The General Assembly
C. The House of Delegates
D. It varies by State
In bicameral State Legislatures, the lower house is called…

A. The House of Representatives  
B. The House of Delegates  
C. The General Assembly  
D. It varies by State

Members of the upper house of State legislatures are elected from districts …

A. Determined by population  
B. Determined by geographical region  
C. That cover the entire State  
D. It varies by State

Which is true of the comparison between the upper and lower houses of state legislatures?

A. The lower house always has more members than the upper house  
B. The upper house always has more members than the lower house  
C. The upper and lower house have the same number of members  
D. It varies by State

The length of terms for members of the upper and lower house in state legislatures is most often…

A. Two years in the lower house and two years in the upper house  
B. Two years in the lower house and four years in the upper house  
C. Four years in the lower house and two years in the upper house  
D. Four years in the lower house and four years in the upper house

The ability to introduce bills is limited to…

A. Only the upper house  
B. Only the lower house  
C. Both houses may introduce bills  
D. It varies by State

Bills introduced to the legislature are discussed in most depth by…

A. Committees that specialize in certain areas  
B. Conference committees that include members of both houses  
C. By the upper house  
D. By the lower house

When the two houses in a bicameral legislature cannot agree on the same version of a bill…

A. A conference committee made up of members of both houses works out a compromise  
B. Each house votes on their own version  
C. The bill immediately dies  
D. The governor chooses the version to support
In a unicameral system…

A. There is only one house of legislature, and all members are equal  
B. There are two houses but they meet together as one  
C. There is only one house, but it contains both Senators and Representatives  
D. There are two houses but all members have equal standing

A unicameral system is used by…

A. No States  
B. One State  
C. Three States  
D. Eight States

Which of the following is not a typical power of governors?

A. Pardoning criminals  
B. Influencing the State budget  
C. Vetoing bills passed by the legislature  
D. Casting the tie breaking vote in the State Legislature

In the United States governors serve for terms of…

A. Four years in all States  
B. Four years in most States, two years in some  
C. Four years in most States, three years in some  
D. Four years in most States, six years in some

In States with the most common form of term limits, after a governor has reached the maximum number of terms they become eligible again…

A. After four years out of office  
B. After eight years out of office  
C. If they have a high approval rating  
D. If the new governor has a low approval rating

Which of the following plans is not used by states with term limits to make a governor eligible to run for office again?

A. Being out of office for four years after serving eight years  
B. Being out of office for four years out of twelve  
C. Being out of office for eight years out of sixteen  
D. Being out of office for eight years after serving four years
In States with term limits for governors, after serving the limited number of terms the governor is excluded from running for…

A. Only the office of governor  
B. All State executive department offices  
C. All State offices  
D. All State and local offices

Which of the following is not a reason for using term limits for governors?

A. Term limits work to prevent the incumbency bias from dominating the office  
B. Term limits work to prevent the consolidation of power in State government  
C. Term limits work to make it easier to remove corrupt governors  
D. Term limits work to make sure the most qualified individuals serves as governor

The incumbency bias refers to the advantage of…

A. Currently serving politicians over a new candidate in elections  
B. Well qualified politicians over new candidates in elections  
C. Corrupt politicians over new candidates in elections  
D. Wealthy politicians over new candidates in elections

Governors are limited to one consecutive term in…

A. No States  
B. One State  
C. Three States  
D. Eight States

It is possible for a governor to be ineligible to run due to term limits even if they never served consecutive terms in…

A. No States  
B. A few States  
C. All States with term limits  
D. Only in States with lifetime term limits

Governors have lifetime term limits in

A. No States  
B. One State  
C. Three States  
D. Eight States
APPENDIX B: Argument Rating Task

The Electoral College is a good system for determining the President of the United States.

The Electoral College is not a good system for determining the President of the United States.

The Electoral College is a good system because it forces candidates to appeal to a wide distribution of the country.

The Electoral College is a good system because it encourages the election of candidates from stable, moderate parties.

The Electoral College is not good system because it does not always follow the popular vote of the nation.

The Electoral College is not a good system because most states use a winner-take-all system.

The Electoral College is a good system because under this system candidates need support from many regions of the country, which is important for ensuring the entire country is represented by the president.

The Electoral College is a good system because it is difficult for radical candidates to get enough support to be elected, and more moderate candidates from stable parties are better able to represent the country as a whole.

The Electoral College is not a good system because under this system it is possible for a candidate who lost the popular vote to become President, and having a President who was elected by the majority of the country would more accurately represent that will of the nation.

The Electoral College is not a good system because the winner-take-all system used in most States limits the effect of many citizens’ votes, and having every citizen's vote count equally would be a more democratic system.

The Electoral College is not a good system, because although the system tries to ensure even small State’s interests are represented, the way it does this causes the value of votes some citizens to be of greater value than others.
The Electoral College is a good system, because even though the value of some individual votes are reduced, the winner-take-all system of most States best represents the will of the population of each State as a whole.

It is better for State Supreme Court Justices to be elected than to be appointed by the Governor.

It is better for State Supreme Court Justices to be appointed by the Governor than to be elected.

It is best for State Supreme Court Justices to be elected because important State officials should be accountable to the State's citizens.

It is best for State Supreme Court Justices to be elected because it is possible that Governors will appoint judges only based on whether they will support their political opinions.

It is best for State Supreme Court Justices to be appointed by the Governor because the commissions used in the Missouri Plan are knowledgeable enough to find the most qualified candidates.

It is best for State Supreme Court Justices to be appointed by the Governor because appointed judges are not influenced by raising money for campaigns.

It is best for State Supreme Court Justices to be elected because judges who are elected are directly accountable to the citizens of the State, and accountable judges will be more concerned about making fair decisions in the interest of citizens.

It is best for State Supreme Court Justices to be elected because Governors may put the political beliefs of candidates above their qualifications, and their political allies may not be the most qualified candidates for the position.

It is best for State Supreme Court Justices to be appointed by the Governor because the people in the nonpartisan commissions that select candidates are more aware of the qualifications of potential candidates than the average voter, and it is important to have the most qualified and deserving candidates in office.

It is best for State Supreme Court Justices to be appointed by the Governor because elected judges must run campaigns and raise donations, and perceived ties to money can reduce the effectiveness of judges by causing citizens to have less respect for their legal authority.
It is best for State Supreme Court Justices to be appointed by the Governor, because while it may be more democratic to elect them, judges are some of the least known officials and the average voter is unlikely to be aware of their qualifications.

It is better for State Supreme Court Justices to be elected, because although commissions may be more aware of the qualifications of candidates than average citizens, the people on the commission may have political agendas with interests that may not line up with the interests of citizens of the State.

A bicameral system of State legislature is better than a unicameral system.

A unicameral system of State legislature is better than a bicameral system.

A bicameral State legislature is best because in a bicameral system the two houses can serve as checks against each other.

A bicameral State legislature is best because the two houses allow greater discussion of bills with more opportunities for compromises.

A unicameral State legislature is best because a unicameral system is better able to efficiently pass legislation.

A unicameral State legislature is best because there is never a situation where a few legislators in a conference committee have the last say over the final version of a bill.

A bicameral State legislature is best because having two houses that serve different terms can represent both long term and short term interests of the State, which is best able to truly represent the will of the State's citizens over time.

A bicameral State legislature is best because having two houses forces the legislature to deliberate bills in more depth as they come to a consensus, and this allows the best version of the bill to be created.

A unicameral State legislature is best because in a unicameral system there is less time and money wasted in coming to a consensus, and it is important to save State resources for where they are needed.
A unicameral State legislature is best because there are no conference committees that have the power to make additions to bills that cannot be further amended, and giving all legislators equal standing on the final bill best represents the citizens.

A bicameral State legislature is best, because although having only one house is able to proceed more efficiently, the extra time spend deliberating bills in coming to a consensus between two houses helps reach the best version of the bill.

A unicameral State legislature is best, because although the final bill resulting from two houses coming to consensus may have been deliberated over more, the efficiency of a unicameral system saves the State time and money that can be scarce resources at the State level.

It is better for Governors to be subject to term limits than be able to continue to be elected.

It is better for Governors to be able to continue to be elected than to be subject to term limits.

It is best for Governors to have term limits because having term limits reduces the effect of the incumbency bias.

It is best for Governors to have term limits because term limits reduce the chances a Governor has at becoming tied to special interests.

It is best for Governors to not have term limits because a Governor can gain in experience as they serve.

It is best for Governors to not have term limits because citizens should have the right to vote for the most popular and qualified candidates, even those already in office.

It is best for Governors to have term limits because term limits lower the disadvantage new candidates face against incumbents, and allowing qualified newcomers to have an equal chance helps ensure the most qualified candidates are elected.

It is best for Governors to have term limits because having term limits makes it more difficult for a Governor to consolidate power, and this lowers the chances for him or her to allow corruption to influence State policy.
It is best for Governors to not have term limits because it is democratic to allow the most popular candidate to be elected by the people, and a candidate with the popularity to be able to elected to office many times may be the best person to represent those citizens.

It is best for Governors to not have term limits because experience can be one of the greatest qualifications for office, and the experience a Governor gains may make him or her better able to serve than a newcomer.

It is best for Governors to not have term limits, because even though term limits can reduce the effects of the incumbency bias, it is likely than an incumbent who is sufficiently popular to be reelected may be the best candidate to represent the will of the citizens.

It is best for Governors to have term limits, because even though Governors may be more experienced after serving in office, they may still run for other State or federal offices and use their experience in a new role while allowing new candidates to serve as Governor.
APPENDIX C: Essay Questions

It is possible for a candidate to win the electoral vote but lose the popular vote. What are some of the consequences of this, and how does it interact with the ideas behind why the Electoral College is used as opposed to the popular vote? Try to support your answer with information from the lessons. Please try to write at least one paragraph.

The candidates that Governors may appoint to courts are often selected by nonpartisan commissions as in the Missouri Plan. What are some of the consequences of using such commissions to choose candidates and how does this interact with the ideas behind appointing judges rather than electing them? Try to support your answer with information from the lessons. Please try to write at least one paragraph.

In a bicameral system there is often the need for the two houses to form a conference committee to come to agreement on different versions of a bill. What are some of the consequences of using conference committees and how do these reasons interact with the ideas behind using a bicameral system rather than a unicameral system? Try to support your answer with information from the lessons. Please try to write at least one paragraph.

In many States with term limits a Governor can become eligible to run for office again after a certain amount of time out of office. What are some of the consequences of this and how does it interact with the ideas behind whether or not governors are subjected to term limits? Try to support your answer with information from the lessons. Please try to write at least one paragraph.
APPENDIX D: Essay Scoring Rubrics

Content (information is present or absent in essay):

Electoral College:

1. The Electoral College is an indirect system of election
2. The number of Electoral votes a state has is determined by state population
3. Candidates win via Electoral vote rather than popular vote
4. Many states use a winner-take-all system to assign electoral votes
5. There are states that use a congressional district method to assign electoral votes
6. There can be / have been instances where a candidate wins the electoral votes but loses the popular vote
7. In some states electoral votes are automatically granted to the winner of the popular vote
8. Any mention of faithless electors and their consequences (faithless electors vote for a candidate other than the one they pledge to vote for, they can be penalized by fines in some states, and they have never affected the outcome of an election)
9. Individual votes in small states are worth more than individual votes in large states
10. The Electoral College overrepresents the worth of votes in small states to protect small state interests
11. Candidates require a wide distribution of support to win the election (they cannot win just a few select populous areas)
12. The outcome of the election in most states is a foregone conclusion (a consequence of the winner-take-all system, or because of solid/red/blue states)
13. The election is mostly decided by swing/purple states that do not always vote for the same party
14. Third party candidates are not realistically viable in the Electoral College
1. In some States judges are selected via general election
2. Elections for judges can be either partisan or nonpartisan
3. Judges who are elected must run campaigns and collect campaign donations like other politicians
4. In some States judges are appointed by the governor
5. Potential candidates for appointment are usually chosen by nonpartisan commission as in the Missouri Plan
6. Appointed judges are often (as in Missouri Plan) eventually subject to a retention vote by the people
7. The Missouri Plan commissions contain lawyers from the state bar association, citizens appointed by the Governor, and is headed by the current chief justice
8. In states that use a modified form of the Missouri Plan (hybrid system states) appointments by the Governor are subject to approval by the legislature or other elected body
9. Judges can be appoint by the state legislature (only in one state – Virginia)
10. Judges are some of the least known political figures and in elections could be voted on party lines or even voted at random; average people may not be able to judge qualifications accurately
11. Commissions in the Missouri Plan are not necessarily nonpartisan and may have a political agenda or bias and choose political allies
12. The interests of the members of the commission do not necessarily line up with the interests of the State’s people (there may be corruption or they may not choose the most qualified candidates)
13. Missouri plan commissions tend to choose from only a small pool of people in the right legal and political circles, leading to a lack of diversity (such as women or minorities)
State Legislatures

1. A bicameral system includes two houses
2. Most states use a bicameral system for their state legislature
3. Members of the lower house come from smaller regions, usually serve shorter terms, and are more numerous than members of the upper house (or members of the upper house come from larger regions, usually serve longer terms, and there are fewer of them than in the lower house)
4. In a bicameral system both houses must pass the same version of a bill for it to become law / must come to agreement
5. When houses cannot agree on the same version of a bill, it goes to a conference committee composed of members of both houses
6. A unicameral system includes only one house
7. Only one state (Nebraska) uses a unicameral system for its state legislature
8. In a unicameral system all legislators have equal standing / equal power
9. A unicameral system can pass legislation in a more timely manner and is more efficient (or a bicameral system is slower and less efficient)
10. A unicameral system wastes fewer resources and is less expensive (or a bicameral system wastes more resources and is more expensive)
11. Bills produced by conference committees cannot be further changed or amended without restarting the entire process
12. Members of conference committees have more power in shaping bills in the committee than other members of the legislature, giving power to only a few individuals
13. The houses in a bicameral system can serve as checks and balances on each other, and a unicameral system does not have these checks and balances
Governors

1. In some States Governors are subject to term limits
2. In some States Governors are not subject to term limits
3. Governor terms are typically four years (two years in some States)
4. In some States with term limits for Governors, the Governor can become eligible again after being out of office for a time
5. In some states with term limits Governors serve two terms (eight years), then must be out of office for one term (four years), and then becomes eligible again
6. In one state (Virginia) Governors are limited to one term (four years), and become eligible again after one term (four years) out of office
7. In some states Governors are limited to being in office for a certain number of years over a period of time (such as 8 out of 12 or 8 out of 16)
8. Some states use absolute term limits, such as after serving two terms (eight years) the Governor is never eligible again
9. Even without term limits Governors have rarely served more than two terms
10. Term limits can prevent experienced Governors from continuing to serve in office
11. Term limits can prevent Governors from solidifying or abusing power, or becoming corrupt
12. Term limits can prevent the incumbency bias from dominating the office and ensure new people or ideas can enter the office
**Argumentative Judgments and Essay Position**

Rate each statement as an argumentative judgment or not. A statement is each complete idea in a sentence (a sentence can contain multiple statements if it contains multiple unique ideas). A statement is an argumentative judgment if it makes a clear judgment about strength, effectiveness, or positivity of one side of a topic (“the Electoral College is the best system for determining the president”). If a statement only contains facts about the topic but does not make a judgment, it is not an argumentative judgment (“the Electoral College is currently used to determine the president”). If overall the statements in an essay are argumentative judgments that support the same side of the topic, then that essay should be rated as taking an argumentative position. If there are argumentative judgments supporting both sides or the essay does not make argumentative statements, then that essay should be rated as not taking an argumentative position.