Evaluating Hypotheses About Active Learning

Leanne C. Powner
LPowner@umich.edu

Michelle G. Allendoerfer
MAllendo@umich.edu

Department of Political Science
University of Michigan
5700 Haven Hall
505 South State Street
Ann Arbor, MI 48109-1045

BIOGRAPHICAL STATEMENTS

Leanne C. Powner is a Ph.D. candidate at the University of Michigan. She can be reached at LPowner@umich.edu.

Michelle G. Allendoerfer is a Ph.D. candidate at the University of Michigan. She can be reached at MAllendo@umich.edu.

ACKNOWLEDGEMENTS

The authors would like to thank Jim Morrow, Neill Mohammad, Andrea Jones-Rooy, Thomas Chadefaux, Papia Debroy, Scott Woltze, Joe Fang Zhou, Shanna Kirschner, and Sarah Croco for assistance in conducting this research, and Bob Mushroe for assistance in data collation. Deborah Meizlish at the University of Michigan's Center for Research on Learning and Teaching was especially helpful in drafting the research design, and Matt Krain and Jeff Bernstein gave helpful comments on earlier drafts. This paper was previously presented at the Annual Convention of the International Studies Association, San Diego, CA, 22-26 March 2006, and the research was performed under University of Michigan IRB-BS Approval HUM00003932. Appendices and replication data are at http://www-personal.umich.edu/~lpowner .

Evaluating Hypotheses About Active Learning

ABSTRACT

We assess the relative effectiveness of two different active learning complements to traditional lecture-based learning. Using a large introductory class at a large public university, we conducted an experiment designed to evaluate whether active learning approaches provide a significant improvement in a student's short-term retention of material over only attending a standard large lecture. In this Introduction to World Politics class, the teaching assistants each taught one section using a common discussion lesson plan and one section using a brief role-play activity. Using multiple regression analysis, we find that the addition of an instructor-led discussion significantly improves student performance on the short answer portion of a brief post-activity assessment, but not on the multiple choice portion. The role-play sections perform significantly better on multiple choice portion than the lecture-only group. In comparing the two treatments to each other, we find no statistically significant difference in the performance of the two groups.

KEY WORDS

Active learning, bureaucratic politics, pedagogy assessment

The last decade has witnessed something of a minor revolution in college teaching practices as instructors, particularly in the 'STEM' disciplines of science, technology, engineering, and mathematics, moved to adopt 'active learning' instructional techniques. These techniques ask students to participate in constructing their own knowledge through discussion, role-play, simulation, problem-based learning, and other methods. Active learning is often contrasted with traditional lecture-based formats, which scholars argue involve mostly passive learning in which knowledge is conveyed directly to the student, with no discovery or processing necessary on the student's part. Students are expected to absorb knowledge given to them by authority figures such as professors (c.f. Bonwell and Eison 1991). Proponents of this approach laud its ability to convey large amounts of information required for success in the field, but opponents of lecture-only education argue that active techniques increase student knowledge, as well as engagement and enjoyment and thus create lifelong learners who enjoy what they do.

Behind this debate, however, lurks the critical but often-unanswered question of, "do active learning techniques work?" Efforts to evaluate the 'effect' of active learning techniques have encountered formidable obstacles, not the least of which is determining what we mean by 'work' or 'effect' (c.f. Prince 2004). Recent efforts in the scholarship of teaching and learning have thus shifted focus toward demonstrating that these newfangled teaching techniques do indeed produce more or deeper learning than traditional instructional methods: greater amounts of information retained, better comprehension, and/or evidence of expanded ability to apply concepts and use higher order thinking skills. Not unsurprisingly, evidence has been mixed, both for the existence and magnitude of any effects at all, as well as what particular types of intervention produce what kinds of effects in what subject domains and environments.[1]

This paper enters the debate by examining the effect of a particular pair of active learning interventions on student learning. Following a lecture on domestic politics explanations for conflict, students in a large introductory world politics class were assigned to one of two treatments: discussion or a brief role-play followed by a brief reflective discussion.[2] We then compare the results of these two groups and a small lecture-only control group on a post-activity quiz to determine if these types of active learning interventions have any measurable impact on student fact recall or analytic comprehension. We hypothesize that both interventions will produce higher student scores than the control group, and the more intensely engaging and interactive role-play with brief discussion format will produce higher scores than the standard discussion. Our particular contribution to this debate is twofold. One, we were able to obtain a comparatively large sample ($n \leq 164$) and a slate of useful control variables to allow us to isolate the effect of the treatment from other potentially confounding factors. Two, we use a blind experimental design to capture the effects of different treatments over a constant base lecture; this allows us to compare the relative effectiveness of the interventions.

We begin with an exposition of the rationale behind active learning approaches and a discussion of the current state of research on the effect of active learning techniques. We then describe our hypotheses, experimental design, subject population, and obstacles to effective experimental research in these settings. Finally, we present and discuss our findings, including a series of difference of means tests and a multivariate analysis. We find that some type of active learning method in addition to a traditional lecture does improve students' mastery of the material in both objective and open-ended evaluation formats. In contrast to our expectations, we do not find that role-playing presents a significant improvement over an instructor-led discussion.

WHAT DO WE KNOW ABOUT THE EFFECT OF ACTIVE LEARNING?

We define active learning techniques as any instructional technique which requires students to apply or process content as part of the learning experience. This includes such diverse approaches as case studies, simulations, service learning, computer-mediated activities, and other forms of experiential education; problem-based learning as is common in the STEM disciplines[3]; the use of film as an application or context of concepts; group projects, in which students learn from and teach each other; and discussion, which may be paired with any of the above or used independently. Lecture, lecture-based recitation questions, and reading assignments are more to less passive, respectively, in their approaches, but still fall short of student intellectual engagement as a primary component of the instructional method.

Conventional wisdom among proponents of active learning approaches is that these techniques improve student learning and are more engaging and enjoyable for the students (c.f. Bonwell 2003; Bonwell and Eison 1991). As Bonwell succinctly states, "in the context of the college classroom, active learning involves students in doing things and thinking about the things they are doing" (Bonwell 2003). Providing students with an opportunity to work through problems and questions on their own is thought to improve on passive learning that occurs in lecture-based classes. The most common active learning technique is student discussion, where students come up with their own answers and work together, though, as Bonwell and Eison (1991) note, even this "technique is not universally admired" (21). McKeachie et al. (1986) argue that discussion is preferable to lecture when the goals are knowledge retention, motivation, and the development of problem-solving and thinking skills.

In a related effort, Brock and Cameron (1999) use Kolb's Experiential Learning Model to

argue that students go through four stages of learning – concrete experience, reflective observation, abstract conceptualization, and active experimentation – and that most students prefer to concentrate on one or two of the stages in their learning. Active learning techniques are uniquely capable of engaging all four stages of learning, therefore appealing to students with a variety of learning styles. Active learning techniques turn students into mentally engaged participants. Finally, proponents of active learning point to the finding that "Students retain 10 percent of what they read, 26 percent of what they hear, 30 percent of what they see, 50 percent of what they see and hear, 70 percent of what they say, and 90 percent of what they say as they do something" (Stice 1987). To varying degrees, active learning techniques are designed to encourage students to both "say" and "say as they do something," thus increasing student learning. As we hypothesize below, we expect any active learning method to improve student learning over a lecture-only style. We further hypothesize that simulations and role-play activities – where students experience a situation and are actively involved in working through the situation – will produce greater improvements in student learning than instructor-led discussion in questions that require higher-order thinking (e.g., essay or short answer questions). Discussion approaches, we hypothesize, will produce greater improvements in multiple-choice assessment instruments, which emphasize recitation of facts.

The vast majority of available studies come from the fields of chemistry and engineering, both of which have identifiable and recognized education subfields. Prince (2004) provides a thorough review of this literature and concludes that findings on the effectiveness of active learning are indeed unstable; results fluctuate across research designs and active learning techniques, with most studies having some notable or serious methodological concern. The number of recent scholarly publications in political science which examine the issue of

effectiveness suggests that the field of political science education is finally beginning to take seriously Rochester's (2003) call for systematic testing of pedagogical proposals. We highlight several recent efforts and their findings, and discuss how our research design helps improve on common concerns with this literature.

Two recent works explore the use of multi-class simulations and their effect on student learning. Daugherty (2003) finds a positive effect for the simulation. Students report, with a two-to-one margin, that they feel they learned more from the simulation exercise than through a standard paper or exam. The use of subjective self-assessments rather than a more objective measure of substantive knowledge leaves the validity of the findings somewhat open to question, though unlike most cases, students were asked to justify why they would prefer or feel they would learn more from a paper or exam. This is useful information for instructor-scholars in their own analysis of their work and in course planning, though its utility for assessing learning from a simulation is low.

Shellman and Turan (2006) also use student-reported gains in knowledge and critical thinking skills in their assessment of a three-day international relations simulation. In an uncontrolled study, some 76.8% of the participating students felt that the simulation enhanced their overall understanding of international relations (2006: 27, Table 2). The authors report no objective assessments, though, so we are unable to determine the true magnitude or statistical significance of the effect.

Both Daugherty (2003) and Shellman and Turan (2006) used multi-day simulations. Recognizing that not all instructors have the ability to devote multiple class days to simulations, particularly when the instructor doubts the simulation's pedagogical usefulness, Baranowski (2006) conducted a quasi-experiment using two sections of introductory American Politics, both

taught by him. One section experienced a brief (single-class-session) simulation of the legislative process; the other did not. Pretest scores were similar for both sections, though we are given no information about the comparability of the two sections in demographic terms. A difference of means test shows a statistically significant gain of about 1.25 points on an 11-point closed-ended posttest (2006: 39-40). Multivariate regression results show, however, that neither prior exposure to the material through high school civics classes, attendance at previous lectures on the material, nor completion of assigned reading had a significant effect on student scores. Though Baranowski does control for grade on the prior exam,[4] he does not control for year in school, gender, major, or other demographic factors that we would also expect to influence scores.[5]

Like many others, Baranowski (2006) conducts his post-test immediately after the simulation activity. The advantage to that is capturing students' memories and experiences while they are still fresh. The disadvantage to that, unfortunately, is that students' memories and experiences are still fresh: they have not yet had time to process their reactions and absorb the lessons to be learned from it. Most are still focused on the short-term mechanics and reactions rather than the deeper elements. In their study of a Congressional simulation, Bernstein and Meizlish (2003) find short-term content-understanding to be virtually identical between the control group and the treatment group taking part in an extensive congressional simulation. However, they do find that long-term (after three years) knowledge retention was notably higher for students that had participated in the simulation. In another evaluation, Meizlish and Bernstein (2003) find that students in the standard lecture course experience greater gains in short-term factual understanding than the simulation classes on 7 of 20 knowledge items, thus casting doubt on the effectiveness of active learning methods in improving factual recall.

The use of control groups is an important way to establish the magnitude or even presence of an effect. In the absence of a control group, pretest methods provide some baseline, but the use of the same instrument as both pretest and posttest introduces an assessment learning effect, and trying to use two different instruments raises the problem of comparability. Carefully designed experiments use randomized or other nonsystematic means to establish control and treatment groups to ensure that the sample populations are comparable. Results can be particularly unreliable especially in cases of self-selection into treatment groups especially.[6]

Krain and Lantis (2006) conduct a two-round experiment to evaluate the Global Problems Summit. The two classes involved had identical reading assignments for two topics, but each had a lecture/discussion on one topic and ran the simulation to address the second topic. Pre-test and post-test scores on a six-item multiple choice assessment were joined by subjective student evaluations of their level of knowledge. Krain and Lantis obtain identical results in both their arms control and convention against torture experiments,. The simulation group achieved statistically significant gains in learning over their pretest scores, but the difference in gains between the experimental and control groups was insignificant for *both* objective and subjective knowledge. Both techniques produced similar amounts of learning. Confidence in the results comes particularly from the paired experimental design; each class served both as control and as experimental group. Even in the absence of information about the comparability of the two classes, the replication of the result in both experiments enhances our ability to accept the results.

One of the best designed studies is the recent work of Lay and Smarick (2006). This study, an excellent example of how instructor-scholars can collaborate to evaluate teaching techniques, avoids the most of the frequent pitfalls mentioned above. It has a large sample size, is simultaneous, does not involve student self-selection, and uses objective indicators of student

learning with adequate controls. It is not without its own substantial concerns, however. The authors have an admirably large sample (posttest n = 149 for the control group and n = 180 for the experimental group) (Lay and Smarick, 2006:139). The large subject pools and the impressive battery of control data collected help them to demonstrate convincingly that the populations are comparable and to obtain statistically reliable results. The authors limit themselves, however, to difference of means tests on a fairly restricted set of outcome indicators. They find a significant difference between their simulation-treated and control groups in 'knowledge of the legislative process,' which is measured by one question, or two if we consider the majority needed to override a veto part of the legislative process.[7] Given the design of the courses, though, with the authors admitting that the treatment group's syllabus included "a strong emphasis" (137) on the legislative process, we are unable to determine conclusively that the gains are from the simulation rather than from the additional lecture emphasis. Additionally, while this experimental design does treat the simulation as complementary to the lecture portion of the course (rather than as a substitute for lecture), the analysis actually compares the effect of lecture-plus-simulation to the effect of lecture-plus-discussion, with no baseline untreated (lecture-only) group.[8]

We attempt to rectify these major issues in our research design. First, we use an objective measure of student learning, in the form of a ten-point quiz (half multiple-choice, half objective free response). This avoids the problems of class aggregates or student-level whole-course data as the assessment of a particular activity, as well as those of using student self-reported knowledge or only limited assessment information. Second, we use a simultaneous, paired research design with no student self-selection.[9] All instructors used centrally generated but randomly assigned lesson plans in their regular sections during a three-day time span, and while

some substantive emphases may have varied slightly across instructors (e.g., choice of examples, student-driven tangents, etc.), casual auditing of instructors as well as data analysis suggest that this did not occur to any extent that may have resulted in bias.[10] Finally, we collect substantial control data to help isolate the effects of simulation and discussion themselves on top of the lecture and in comparison to one another. This allows us to use multivariate analysis to examine the effect of each treatment and also to compare the effects of different active-learning treatments.[11]

SUBJECT POOL, RESEARCH DESIGN, AND HYPOTHESES

Introduction to World Politics regularly covers domestic explanations for conflict. This provided the context for us to use an activity, previously developed by Powner and Croco (2005) as our treatment of interest. In the Winter 2006 term, the course was taught by Professor James Morrow. Fourteen discussion sections contained 299 undergraduate students, with each section under the subordinate instruction of one of seven teaching assistants. Students hear two fifty-minute lectures from Professor Morrow each week, and meet with their discussion instructor in groups of up to 25 students for two additional 50-minute sessions each week. Enrollment includes students at all undergraduate educational levels (freshman, sophomore, etc.); students are drawn primarily from the University of Michigan's College of Literature, Science and the Arts (LSA), though a small minority of students from other colleges are enrolled as well. Additionally, one section is designated as an 'honors' section, and all enrollees are drawn from students in the LSA Honors Program.

Because each teaching assistant, known as a Graduate Student Instructor (GSI) in University of Michigan parlance, leads two discussion sections, we opted for a matching design

in which each GSI taught a common discussion-based lesson plan in one randomly selected section, and then led the role-play activity in his or her other section. This allows us to hold instructor constant and vary the lesson plan to ensure that differences between instructors are not driving the results. The combination of matching and random assignment helps to reduce the effect of instructor characteristics as a potentially confounding factor, though we acknowledge that we cannot entirely eliminate this from the analysis.[12] In addition, two sections served as our control group and did not receive either treatment – discussion or simulation – prior to students' assessment. The main lecture was held on February 8, 2006; discussion sessions were held over three days, February 8-10, 2006.[13]

Because of the nature of classroom interaction in both discussion and role-play, our common lesson plans contain only approximate directions for instructors. The discussion lesson plan suggests several question prompts and indicates key points that instructors should attempt to cover. The role-play lesson plan includes primarily step-by-step instructions for how the simulation works, but the actual content of the simulation is driven entirely by student reactions and contributions.[14] At the end of each lesson plan, instructors were asked to give students ten minutes to take a brief assessment (See Appendix). This consisted of five multiple choice questions and a brief 'short answer' question of the format used on the examinations. The multiple choice questions were intended primarily as a "factual" recall check; the short answer question, on the other hand, required a deeper understanding of the material. The third page of the assessment included questions about whether the student attended lecture and/or section for that class day, whether the student has done the assigned reading on that material, and the student's class level.[15] To encourage honest responses and to ensure scoring consistency, all scoring was done by the authors rather than by students' own GSIs. The third page of personal

information was also removed from the assessment document before scored assessments were returned to the students' instructors. While the assessment does indicate that it is to "evaluate the effectiveness of class activities," students were unaware during the class period that they were participating in a research project.

To create a control group, two discussion sections led by the first author, including the Honors section, took the assessment at the start of the section meeting, prior to experiencing any other review of the target material. [16] This created a baseline group of student scores which reflect experiencing the lecture only.[17]

Before the assessments were returned to students, GSIs read and distributed a debriefing document to students explaining the nature and purpose of the activity in which they had participated. This debriefing also explained why students were not informed that they were participating in research prior to the activity itself. Students were then invited to read an informed consent form and ask questions, and were repeatedly told that they could choose to sign (or decline) with no penalty or grade influence. The analyses presented here only include the scores of students who chose to sign the informed consent document. Of the approximately 250 students eligible to participate in the study, approximately 175 chose to sign.[18] Our final sample contains at maximum 164 observations.[19]

In line with the general sense of the literature, we hypothesize that students exposed to either of the treatments will score higher on the assessment, *ceteris paribus*, than those in the lecture-only control group.[20] Furthermore, we hypothesize that the students experiencing the role-play treatment will experience larger gains than the discussion-based treatment group. The discussion treatment group's gains should be concentrated in the multiple-choice portion of the assessment, which tests the retention of facts. The role-play treatment group should see gains in

the short answer portion of the assessment, which requires integrative thinking rather than simply factual recall. We note that we are only testing the effects of active learning techniques on knowledge acquisition, rather than student enjoyment. Additionally, due to time constraints, we were only able to test short-term knowledge recall. Ideally, we would like to follow Bernstein and Meizlish (2003) and test long-term recall, but as we did not control the content and timing of examinations or have follow-up opportunities, we must use short-term assessment as an interim step. To summarize, our main expectations are:

1. Students participating in either discussion or role-play activities will score higher overall than non-participants, all else equal.
2. Students who participate in role-play will score higher on the short answer question than the discussion or control groups, all else equal.
3. Students who participate in the discussion will score higher on the multiple-choice questions than the role-play or control groups, all else equal.

ANALYSIS AND DISCUSSION

As a first look at the data, we begin with a series of difference of means tests. Throughout the results section, we will analyze three different dependent variables: total quiz score, multiple choice score, and short answer score. The multiple choice and short answer portions of the assessment are quite different assessment instruments, one tapping fact recall (multiple choice) and one tapping the student's ability to think critically by assembling disparate parts of the material into a single answer. The total quiz score may thus mask important differences in the relationship between the treatments and the student's performance on the quiz. Including both

forms of assessment in the quiz and then disaggregating the scores thus allows us to examine more nuanced hypotheses about the effects of active learning than would otherwise be possible.

[Table 1 about here]

The first two lines of Table 1 compare the lecture-only control group to the treated groups, the traditional discussion format and role-play activity. These difference of means tests show that students that attend a traditional discussion section perform better than students who only attend lecture. On the total quiz score, students in the discussion group earned, on average, 0.776 points higher (out of 10) than students who only attended lecture. Most of this difference appears to be in the short answer portion of the assessment, where students in the discussion group received 0.704 points higher (out of 5) than their lecture-only counterparts in the control group. Both of these results are statistically significant at least at the 0.10 level; the short answer result is statistically significant at the 0.05 level. The difference of performance on the multiple choice portion is neither substantively significant (0.07 out of 5), nor statistically significant.

Next, we consider the difference of means for the lecture-only control group and the students who participated in the role-play activity. Only in the case of the short answer portion did the role-play group perform better than the lecture-only group (0.151 out of 5 points). In the case of all three assessment measures, the difference of means between the lecture-only and the role-play groups are not statistically significant.

Finally, we consider the difference of means for the two treated groups: the traditional discussion groups and the role-play group. The difference of means for the total quiz and short answer portion of the quiz are both statistically significant at the 0.05 level with the discussion groups performing slightly better than the role-play groups. The discussion groups earned, on average, 0.83 points higher on the total score and half a point higher for the short answer portion.

In addition to the difference of means analysis, we also used multivariate analysis and controlled for other variables we expect would influence a student's performance on the assessment. These data are compiled from self-reported information on a separate page of the assessment which was removed before the student's own GSI saw the quiz results. The regression results presented in Tables 2-4 include controls for: the experience of the student's GSI[21], the number of days between lecture and the assessment, whether the student had completed the reading before class, whether the student played the role of the President or a cabinet member in the activity, and class year.[22]

[Table 2 about here]

Table 2 compares the lecture-only control group to the students participating in the traditional discussion section in addition to attending lecture. The results for the treatment explanatory variable are statistically significant and positive in two of the three cases. The only exception is when the dependent variable is the multiple choice portion of the assessment. Here the coefficient on the treatment variable is still positive, but not statistically significant. Substantively, the coefficient in the total quiz score regression suggests that a student who attends lecture and then participates in a GSI-led discussion section will earn 1.9 more points (out of 10) on the quiz than a student who only attends lecture, holding everything else constant; nearly all of this improvement is in the short answer score. A student in the discussion section earns 1.885 points higher on the short answer portion of the assessment than a student who only attends lecture.

[Table 3 about here]

Table 3 presents the results for the second set of regressions which compare the lecture-only control group to the students who participated in the role-play activity. The results for all

three dependent variables are positive, suggesting that the students in the latter group perform better than the students who only attend lecture. A student in the role-play group would earn, *ceteris paribus*, 2.4 points higher (out of 10) on the total quiz than a student attending only lecture. In this comparison, performance on the multiple choice portion improves for students participating in the role-play activity. The results for total quiz score and multiple choice score are statistically significant. The results for the short answer portion are not statistically significant. These results suggest that the role-play improves multiple choice performance, but not short answer scores. This result, however, may be an artifact of the common discussion plan and the short answer question on the assessment; although written separately, the short answer question corresponded tightly to the discussion plan.

[Table 4 about here]

The final set of regressions compares the two treatment groups to each other. These results are presented in Table 4. The students treated to the traditional discussion section are coded as a zero, while the students participating in the role-play are coded as a 1. The results on the treatment dummy never approach statistical significance. In all three cases, the coefficient on the treatment variable is negative, suggesting that the discussion section groups perform better than the role-play groups. This result may reflect the greater ability of instructors to tailor or focus discussion material on assessed information as well as to cover that material in greater detail. However, as these results are not statistically significant, we can only conclude that there is no difference between these two active learning techniques in our analysis.

While non-findings in particular are not often of interest, the substantive implications for pedagogy here are clear. Parallel with Krain and Lantis (2006), we find that engagement of any variety – discussion or role-play or simulation – increases student performance above lecture-

only instructional techniques. The various techniques, however, produce no discernable differentiated effect on student learning; all are equally good for improving learning, at least in the short to medium term. Since some types of active learning require substantially more, and more intensive, preparation on the instructor's part, this suggests that balancing the time demands of preparation with competing demands has no major effect on student learning. While we do not take this to mean that simulations should be eliminated in favor of less time-consuming pedagogical techniques, we do encourage users to think carefully about the pedagogical goals of any activity before incorporating it into the syllabus (Kille et al. 2007). Simulations (or other intensive activities) solely for activity's sake are not beneficial for students or instructors.

CONCLUSIONS

We have several sets of findings. One is that, as expected, either treatment improves performance over lecture alone, holding everything else constant; role play improves total quiz scores by about 2.4 points and discussion by 1.9 (both $p < 0.5$).[23] The effects of each treatment, however, are different. Role play boosts performance in the multiple-choice section of the assessment, and discussion in the short-answer portion. This counters theoretical predictions, which expect that performance on analytical tasks like open-ended response items would improve from the more engaging treatment (simulation).

Comparing the two active learning techniques head to head, though, reveals that role-play does not produce significantly larger gains than discussion. The coefficient on the simulation variable is statistically insignificant, though we do note that the sign is contrary to our theoretical expectations. This finding holds even when controlling for major alternative explanations like

class year, days since the lecture, intensity of participation in the simulation, and the instructor's teaching experience. We can conclude, then, that both techniques are similarly effective; neither is superior to the other for the types of tasks assessed here.

That discussion sections add value to a course is probably not surprising; that is, after all, their primary purpose – not just providing employment for graduate students. Many instructors and scholars are more interested, though, in other types of active-learning methods, such as the role-play activities and simulations. One common concern with integrating role-playing activities into classrooms is that these methods are less efficient, preventing instructors from covering as much material in a given time. Our results suggest that this concern may be unwarranted. Although the groups that participated in the role-play activity did not perform statistically better than the groups that were engaged in a traditional discussion section, the role-play groups also did not perform significantly worse in two out of three evaluations (total quiz score and multiple choice score). The lack of significance may be due to the small size and particular nature of the control group as much as it is to the effect of the activity itself.

Activities such as the bureaucratic politics role-playing game may serve to engage students whose learning styles benefit from more active approaches to learning. Unfortunately, we were not able to evaluate the hypothesis that students with different learning styles benefit differently from the two treatments in this experiment. Another hypothesis we were unable to evaluate here is whether active learning approaches, such as simulations, produce a more substantial benefit in the long-term than in short-term. As Meizlish and Bernstein (2003) find, the benefits of active learning methods may be more pronounced in long-term knowledge retention than is demonstrated in a short-term recall instrument as we used. These unanswered questions provide interesting direction for future research.

REFERENCES

Baranowski, Michael. (2006). Single Session Simulations: The Effectiveness of Short Congressional Simulations in Introductory American Government Classes. *Journal of Political Science Education* 2: 33-51.

Bernstein, Jeffrey L., and Deborah S. Meizlish. (2003). Becoming Congress: A longitudinal study of the civic engagement implications of a classroom simulation. *Simulation and Gaming* 34: 198-219.

Bonwell, Charles C. (2003). *Active Learning: Creating Excitement in the Classroom*: Active Learning Workshops, Green Mountain Falls, CO: electronic file available at http://www.macomb.cc.mi.us/arc/DLmaterials/Bonwell.Active%20Learning.January%202003.pdf. Accessed 12 March 2006.

Bonwell, Charles C. and James A. Eison. (1991). *Active Learning: Creating Excitement in the Classroom.* ASHE-ERIC Higher Education Report 1.

Brock, K., and B. Cameron. (1999) Enlivening Political Science Courses with Kolb's Learning Preference Model. *PS: Political Science and Politics* 25(3):251–256.

Daugherty, Beth K. (2003). Byzantine Politics: Using Simulations to Make Sense of the Middle East. *PS: Political Science and Politics* 36: 239-44.

Frederking, Brian. (2005). Simulations and Student Learning. *Journal of Political Science Education* 1: 385-94.

Krain, Matthew, and Christina J. Shadle. (2006). Starving for Knowledge: An Active Learning Approach to Teaching About World Hunger. *International Studies Perspectives* 7(1): 51-66.

Krain, Matthew, and Jeffrey Lantis (2006). Building Knowledge? Evaluating the Effectiveness of the Global Problems Simulation. *International Studies Perspectives* 7(4): 395-407.

Lay, J. Celeste, and Kathleen J. Smarick. (2006). Simulating a Senate Office: The Impact on Student Knowledge and Attitudes. *Journal of Political Science Education* 2(2) (May): 131-146.

McKeachie, Wilbert J., Paul R. Pintrich, Yi-Guang Lin, and David A.F. Smith. (1986). *Teaching and Learning in the College Classroom: A Review of the Research Literature* . Ann Arbor: Regents of the University of Michigan.

Meizlish, Deborah S., and Jeffrey L. Bernstein. (2003). Unpacking the 'Education' in Civic Education. Paper presented at the International Civic Education Conference, New Orleans, LA, 16-18 November 2003.

Powner, Leanne C., and Sarah E. Croco. (2005). Making Formal Models Freshman-Friendly. Paper presented at the International Studies Association Annual Convention, Honolulu, HI, 1-5 March 2005.

Prince, Michael. (2004). Does Active Learning Work? A Review of the Literature*, Journal of Engineering Education*: 223-231.

Rochester, J. Martin. (2003). The Potential Perils of Pack Pedagogy, or Why International Studies Educators Should Be Gun-Shy of Adopting Active and Cooperative Learning Strategies. *International Studies Perspectives* 4(1): inside back cover.

Shellman, Stephen M., and Kürşad Turan. (2006). Do Simulations Enhance Student Learning? An Empirical Evaluation of an IR Simulation. *Journal of Political Science Education* 2: 19-32.

Stice, James E. (1987). Using Kolb's Learning Cycle to Improve Student Learning. *Engineering Education* 77(5): 291-96.

Table 1: Difference of Means

	Total Quiz Score	Multiple Choice	Short Answer
Lecture-only to Discussion	-0.776* (0.456)	-0.072 (0.242)	-0.704** (0.298)
Lecture-only to Role-play	0.05 (.467)	0.205 (0.241)	-0.151 (0.316)
Discussion to Role-Play	0.830** (0.350)	0.277 (0.181)	0.55** (0.28)

* $p < .10$ ** $p < 0.05$

Table 2: Regression Results – Comparing Lecture-only to Lecture plus Discussion

	Total Quiz Score	Multiple Choice	Short Answer
Treatment	1.919*	0.034	1.885**
	(1.036)	(0.559)	(0.678)
Days since Lecture	-0.148	-0.152	0.003
	(0.216)	(0.116)	(0.141)
TA Experience	0.128	-0.054	0.181
	(0.206)	(0.111)	(0.135)
Reading	0.307	-0.139	0.446
	(0.421)	(0.227)	(0.275)
Class Year	-0.528**	-0.233*	-0.295*
	(0.236)	(0.127)	(0.154)
R^2	0.13	0.06	0.18
No. Observation	94	94	94

* $p < .10$ ** $p < 0.05$

Table 3: Regression Results – Comparing Lecture-only to Lecture plus Role-play

	Total Quiz Score	Multiple Choice	Short Answer
Treatment	2.397**	1.577**	0.821
	(1.201)	(0.599)	(0.860)
Days since Lecture	-0.111	-0.092	-0.019
	(0.199)	(0.099)	(0.142)
TA Experience	0.507**	0.410**	0.097
	(0.248)	(0.123)	(0.177)
Reading	0.431	0.081	0.350
	(0.422)	(0.220)	(0.317)
Pres/Cabinet Member	0.037	0.259	-0.222
	(0.550)	(0.274)	(0.394)
Class Year	0.756**	0.293*	0.462*
	(0.335)	(0.167)	(0.240)
R^2	0.12	0.17	0.081
No. Observation	84	84	84

* $p < .10$ ** $p < 0.05$

Table 4: Regression Results – Comparing Discussion to Role-Play

	Total Quiz Score	Multiple Choice	Short Answer
Treatment	-0.587	-0.172	-0.415
	(0.481)	(0.248)	(0.332)
Days since Lecture	0.063	-0.077	0.139
	(0.236)	(0.121)	(0.163)
TA Experience	0.333**	0.145*	0.188*
	(0.656)	(0.084)	(0.112)
Reading	0.500	-0.215	0.715**
	(0.400)	(0.206)	(0.276)
Pres/Cabinet Member	-0.645	-0.137	-0.508
	(0.517)	(0.266)	(0.357)
Class Year	-0.207	-0.077	-0.130
	(0.216)	(0.111)	(0.149)
R^2	0.12	0.07	0.16
No. Observation	120	120	120

* $p < .10$ ** $p < 0.05$

APPENDIX. Activity Assessment.

PS 160 Intro to World Politics
Bureaucratic Politics Activity Assessment

Name _____ Sec _____

The results of this assessment will help the PS 160 GSIs to determine which class activities are most effective for conveying information. Curious students are welcome to discuss the results with Leanne Powner after mid-March. Please answer to the best of your ability. *The assessment continues on the back and on the second page.*

I. Multiple Choice. Print the letter of the best answer on the line at left.

1. _____ According to Allison, 'chiefs' are _____.
   A) permanent civil servants who make daring or risky policy proposals
   B) Cabinet members and other prominent Administration figures
   C) bureaucrats who head the State Department's regional bureaus
   D) prominent opinion leaders outside government.

2. _____ SOPs do all of the following *except* _____.
   A) assist large groups of actors to coordinate their actions
   B) reduce the time needed to respond to a crisis
   C) consistently produce the best results for each individual crisis
   D) result in incremental organizational changes

3. _____ Political appointees and staffers refrain from showing doubt when advising the president. We refer to this as _____.
   A) the 51-49 principle
   B) the doctrine of "where you sit depends on where you stand"
   C) standard operating procedures
   D) internal politics

4. _____ According to the government politics perspective, the president might pick a particular policy because _____.
   A) he trusts that policy's proposer more
   B) that policy's proposer argued in favor of his proposal more successfully than others argued for theirs
   C) that policy's proposer had superior access to the President
   D) all of the above

5. _____ All of the following are common critiques of the bureaucratic politics perspective, *except* _____.
   A) it is difficult to apply to other countries
   B) it views states as unitary actors
   C) it ignores international pressures
   D) it ignores other avenues of domestic pressures

II. Short Answer. Answer the following question in 2-4 sentences, as if this were a short answer question on an exam.

> How do organizational processes approaches explain the president's decision-making process? How do governmental politics approaches explain the president's decision-making process? In your answer be sure to identify the actors in each approach, what those actors do, and how this affects presidential decision-making.

CONTINUES ON NEXT PAGE → → →

Name _____ Sec _____

III. True/False. ***Non-Scored.*** Answer true (T) or false (F) to each of the following statements. The answers to this section are *not* part of the quiz score and this page will be removed before your GSI sees the scores. Please answer honestly.

T F I attended Prof. Morrow's lecture on Wednesday, February 8, about domestic explanations for conflict.

T F I completed the assigned reading for the bureaucratic politics class (short section of the Bueno de Mesquita text, Allison article on the Cuban Missile Crisis in the course pack).

T F I attended my discussion section on the class where we covered bureaucratic explanations for conflict.

T F I played the role of a Cabinet member or the President in the simulation today. * If this is not applicable to your section, mark "F."

Which of the following best represents your level in college? (circle one)

Freshman Sophomore Junior Senior

_____

For Scorer Use Only.

RIN	D	G	MC	SA

[1] Prince (2004) contains a comprehensive review of such evaluations in the field of engineering education; we review prominent contributions to the political science education literature below.

[2] See the 'Bureaucratic Politics Game,' in Powner and Croco (2005). In addition, a small control group was retained; see discussion below.

[3] See Prince 2004 for a discussion.

[4] This is an important effect to consider. We are unable to include it in our own study, however, as no exams or papers had yet occurred in the class and no other course-wide assignment was available for substitute.

[5] Baranowski (Appendix A, pg. 45) does collect data on class year, previous study, gender, etc., but does not include them in the reported multivariate results. Note 3 (pg. 42) does suggest, however, that unpublished models including previous study produced insignificant results.

[6] Krain and Shadle (2006), for example, conclude that participants in the Hunger Banquet role-playing activity "demonstrated a greater degree of knowledge acquisition than students who learned the same material in a traditional classroom setting" (52) Their treatment group, however, is a self-selected sample. Undergraduates who are willing to give up their time, even with the lure of free food, are probably more motivated and paying more attention than the average class, even if the class is substantively relevant and composed of upper-division students. Krain assures us, however, that tests show no statistically significant differences exist across the groups on key variables (personal communication, 28 March 2006).

[7] With such large and comparable subject groups, using an assessment instrument focused around the specific learning goals of the simulation would have provided a strong opportunity to pool the subjects and estimate the effect of the simulation itself on student knowledge of the

legislative process. This is particularly true if the control group had been given a similar amount of emphasis on the legislative process.

[8] Frederking (2005) finds that classes which included a congressional simulation preformed better on assessments than did classes not participating in the simulation. However, the three examinations used as assessment tools occurred prior to the actual simulation (Frederking, personal communication, May 24, 2006), indicating that the improvement could not be attributed the simulation itself.

[9] Some selection may occur via the consent process, but as we discuss below, given the number of students involved this seems to be only a marginal concern.

[10] Discussions with the instructors revealed that some had used different examples (Hurricane Katrina, 9/11, university processes, or student-suggested) to illustrate the limitations of standard operating procedures in their discussion sections. Simulation groups, as can be expected with these types of activities, had differing experiences of topics and actions as a result of their student-driven nature. No questions on the assessment asked explicitly about the examples that instructors reported discussing. Auxiliary models not reported here controlling for instructor (instead of instructor experience, as reported here) reveal no systematic effect beyond that captured by instructor experience, so we are reasonably confident that no serious bias occurred.

[11] Course design did not allow us to administer the posttest at a later date. Because sections meet twice a week (Monday-Wednesday or Tuesday-Thursday) in most cases but only once a week for the Friday sessions, the time between treatment and posttest would have varied from three days to seven days, and the time from lecture to posttest would have varied from five days to seven days. Even controlling for this, such a design would have risked conflating the effects of

the treatment with any additional reading, discussion, review, or mental processing that the students did between activity and assessment.

[12] The analyses that follow control for GSI experience as an additional means of reducing this effect.

[13] Multivariate analyses below control for the number of days between lecture and section to accommodate the possibility that students reviewed their notes, had additional time to complete the readings or generate questions, etc. This variable is consistently insignificant by conventional measures, much to our dismay as instructors.

[14] Lesson plans for both the discussion and the role-play activity are available from the first author's web site, http://www-personal.umich.edu/~lpowner .

[15] We opted for self-reported class level rather than the official level recorded by the registrar because this allowed the student, whom we assume is more familiar with his/her situation than we are, to determine the appropriate level. Many first-year students, for example, have sophomore standing as a result of AP credits, but they self-identify as first years. Some students in their third year are on an accelerated schedule as a result of AP credits or summer study and see themselves as in the same position as seniors.

[16] The inclusion of the honors section in the control group may create some bias in the control group. We evaluated this possibility and found that, using a simple difference of means test, the scores between the honors control group and the other control group was not statistically significant at conventional levels. Additionally, we repeated the multivariate analyses with a control variable for the Honors section, but this did not affect the results. The consistently insignificant coefficients on this variable, however, led to its eventual exclusion from the final analyses. These analyses are available from the authors by request.

[17] Allendoerfer scored Powner's classes' assessments to encourage honest reporting; students were informed of this prior to the administration of the assessment. These completed assessments were treated in the same manner as student evaluations – a student took them directly to the department office and gave them to Allendoerfer.

[18] A small number of students whose consent forms were signed did not have assessment scores on file; these students were excluded from the study despite their consent. Additionally, because of time constraints caused by a cancelled class meeting, one GSI declined to participate in the study.

[19] Because we are interested in the impact of the treatments in addition to lecture, students who reported that they did not attend lecture we excluded from our analysis. This reduces our sample size for 175 consenting students to 164.

[20] An unintended consequence of our matching design was that our control group was much smaller than either treatment groups. We recognize that, ideally, our control group would be of comparable size to the treatment groups.

[21] We measured GSI experience as the number of terms the GSI had taught this course. As the same text and essentially the same syllabus have been in use for five years, we felt comfortable using previous teaching experience in this course as an indication of both comfort with the material and with appropriate pedagogical strategies for it.

[22] We had also asked students to report if they attended lecture and section. We did not include these control variables because all students reported attending section and nearly all reported attending lecture. Because we asked the GSIs to administer the assessment on the same day as the activity, it is not surprising that all students reported attending section on the day of the activity/discussion; students who missed that section meeting (an admittedly nonrandomly

selected group) were effectively eliminated from the sample. Only 12 of 164 students reported not attending lecture. This is somewhat surprising because the students had a paper due in lecture the day this topic was covered; in previous terms, students who pulled all-nighters on the paper frequently turned in their papers and then left again without staying for the lecture.

[23] These results are robust when controlling for Honors status. In our analyses, students in the Honors section never performed significantly better than other students.