Making Assignments Count and Other Strategies for Increasing Critical Thinking in Political Theory

Journal:	*Journal of Political Science Education*
Manuscript ID	JPSE-2017-0036.R2
Manuscript Type:	Scholarship of Teaching and Learning
Keywords:	Critical thinking, political theory, Undergraduate education, domain specificity, Assessment

Making Assignments Count and Other Strategies for Increasing Critical Thinking in Political Theory

Abstract

Political theory lags behind other subfields in political science in rigorously testing what helps foster critical thinking (CT). Yet some of the greatest temptations to engage in motivated reasoning can be found in normative political contexts. This study uses multiple regression analysis to explore 9 semesters of data from an introductory course in political theory. Three results stand out. 1. As the stakes of an assignment decrease, so do CT scores. 2. As the number of course assignments increase, CT scores fall. 3. When preparatory exercises matter, theory-specific CT exercises outperform generic ones. A theory of student rationality is put forth. Implications for course design, program assessment, and future research are discussed.

Introduction

Colleges and universities in the United States often see themselves as places where students develop their critical thinking (CT) skills[1]. In political science, multiple articles have outlined and evaluated strategies for promoting CT in the classroom. These include: opportunities for critical self-reflection through writing (Cavdar and Doe 2012), engaging with social science methodology (Marks 2008, Olsen and Statham 2005), group discussion (Blings and Maxey 2016, Williams and Lahman 2011), group debate (Omelicheva 2007, Oros 2006), and the avoidance of political controversy (Fitzgerald and Baird 2011).

[1] According to a report from the Association of American Colleges and Universities, the ability for critical analysis and logical thinking is mentioned in fully 25% (75 of 301) of mission statements of the Princeton Review's "331 Best Colleges and Universities in the United States" that express an educational vision. This places critical thinking in fifth place on the list of most common educational themes in university mission statements.(Gaff and Meacham 2014)

Although introductory courses, American government courses, and courses in comparative politics are well-represented in this literature, strategies for developing CT in political theory have not been empirically assessed. Given evidence that CT may have important subject specific components (McPeck 1981, Smith 2002) and moreover that moral reasoning skills probably differ from other kinds of reasoning skills (Mason 2007, Phillips and McMillian 2010) a gap in the literature exists with respect to what promotes CT in the subfield of political theory. By leveraging data collected from 451 essays over the course of 9 semesters in a required junior-level course in political theory, this paper begins to fill this gap.

Why Critical Thinking?

The ability to think critically, that is to say the ability of individual citizens to construct and evaluate conclusions from available evidence and assumptions (Williams and Worth 2001), is a *sine qua non* of a functioning democracy and economy. The idea that democracy fares roughly as well as the ability of its citizens to critically evaluate the rhetoric of its leaders is one that can be traced as far back as Plato. The idea that analytical skill (a component of CT) is important to the job performance of employees is also well-cemented in today's economy, including among American employers (NACE Job Outlook, 2016)

The role of colleges and universities in preparing individuals for their roles in these contexts is hardly disputed either. Faculty have long rated CT among their most cherished pedagogical objectives (Bok 2006; Paul 1991; Resnick and Peterson 1991) and universities around the country have focused their attention on developing CT for some time now.

Research projects, such as project CAT (Stein et. al 2010), the California Critical Thinking Disposition Inventory (Facione, Facione, and Sanchez 1994), or the Washington State University Critical Thinking Project (Kelly-Riley. et. al. 2008) have developed assessment tools to try and measure CT. Exams like the California Critical Thinking Skills Test (CCST) and the Collegiate Learning Assessment (CLA) are routinely administered by universities to check for progress in CT in their graduates (though not without controversy: see Klein, Shavelson, and Bolus 2007).

Political theorists also want to develop CT skills in their students. In Matthew Moore's 2011 survey "How (and What) Political Theorists Teach: Results of a National Survey," fully 96.3% of political theorists surveyed thought that inculcating CT skills in their courses was either "important" or "very important." This is unsurprising since a wide range of viewpoints in political theory agree on the importance of CT.

Critical theorists, Feminists, and Marxists want to peel back the illusions of social norms, power, gender, race, and class. To do this, practitioners must be capable of critically reflecting on the assumptions, categories, and processes that characterize the modern

world. Liberals and perfectionists cherish CT because of its connection to autonomy: the capacity of individuals to set ends for themselves. On these views, the twin-projects of democracy and individual liberty remain radically incomplete if citizen choices are the product of unquestioning acceptance of existing cultural norms or power structures.

Operationalizing Critical Thinking

Despite near-universal agreement on its importance, there are competing definitions of CT. Pascarella and Terenzini (2005) analyze different conceptualizations and find the following broad regions of agreement. "Most attempts to define and measure critical thinking operationally focus on an individual's capability to do some or all of the following: identify central issues and assumptions in an argument, recognize important relationships, make correct references from the data, deduce conclusions from information or data provided, interpret whether conclusions are warranted based on given data, evaluate evidence of authority, make self-corrections, and solve problems." (156)

CT in this study was measured using a modified version of the Washington State University critical thinking rubric. Over a period of 9 semesters, 451 1,000-1,500 word essays were assessed in a required political theory course aimed at junior political science majors at a midsize southern public university.

Six categories, broadly reflective of the consensus articulated above, are independently scored. The Focus score evaluates how well a student conceptualized the question being asked. The Interpretation score looks at how fairly students interpreted their sources in those essays. The Thesis score assesses the logical consistency of the argument presented. The Objections score reflects how well a student understood and responded to the arguments raised by alternative perspectives. The Evidence score measures whether appropriate sources, facts, and normative principles were marshalled in support of the thesis. Finally, the Conclusion score measures the ability to assess the significance and implications of the argument made (for public policy, political theory, and so on)[2].

The choice was made to report results for each dimension of CT separately as well as an aggregate measure. There is no a priori reason to believe that the interventions or course designs that (for example) help students better select evidence for their argument also help them with their logical reasoning or their ability to interpret the meaning of the question. My experience using this grading rubric is also that students may write very cogent papers (high Thesis scores) that are completely off topic (low Focus scores). They may develop their perspective very well (with respect to logical consistency and evidence) and yet ignore obvious objections. This variation in subcategory scores seems worth preserving.

[2] A more detailed discussion of the WSU grading rubric and adaptations made for political theory is available in Appendix A.

Several checks on the reliability of the CT measures were conducted as a post-test using a random sample of 24 essays scored by 3 different graders. A summary of the reliability checks is available in Appendix B.

Stimulating Critical Thinking

The data set permits us to test the relative effectiveness of several different types of exercises for stimulating CT including critical review essays (CREs), critical thinking exercises (CTEs), and logic game exercises (LGEs).

Logic Games

An important component of CT is the ability to draw accurate inferences from available information. Plausible induction and valid deduction are central to the application of CT to any given context. In that sense, logic is a kind of "generic" CT skill, since the rules of formal logic can be applied successfully in any context. A popular textbook on logic claims, "the study of logic is one of the best ways to increase students' skill in critical thinking" and "The study of logic increases one's ability to understand, analyze, evaluate, and construct arguments." (Howard-Snyder, Howard-Snyder, and Wasserman 2012: xv)

Because of the question of domain specificity raised earlier, whether CT skills (or any kind of reasoning skills) are easily transferred from one field of inquiry to another is still

a topic of debate. On the transferability side of the debate see Ennis (1989) and Bailin (2002). On the "domain specificity" side see Halpern (2001) and Van Gelder (2005).

During three semesters, students practiced LGEs of the kind one finds in puzzle books or the Law School Admissions Test (LSAT) every two weeks for one class period. At the beginning of the semester, students were subjected to a short introduction to formal logic. Before the first LGE, students were walked through some examples of logic puzzles and common types of operations seen in those puzzles (particularly *modus ponens* and *modus tollens*) After most students had solved the games, the class as a whole was asked to assess of the potential soundness of the conclusions reached.

Though LGEs incorporated themes from political theory, they did not provide opportunities for moral reasoning but instead consisted of a series of deductive logical operations that led to a unique set of possible conclusions. As such, they are very unlike the more nuanced, probabilistic, self-aware, and contextual kind of reasoning characteristic of high quality essays in political theory. Still, if CT is a generic skill that can be both practiced and applied to political theory, then LGEs should help with the rigor of the essays if nothing else[3].

Critical Review Essays and Critical Thinking Exercises

Two other interventions stand out as domain-specific counterparts to the attempt to cultivate generic CT skills: critical review essays (CREs) and critical thinking exercises

[3] The thesis score stands out as the most likely candidate for improvement.

(CTEs). For the CREs, each student received an anonymous essay written by a classmate that semester and was tasked to write a review essay outlining its strengths and weaknesses.

Though not directly an opportunity for self-reflection as suggested by Çavdar and Doe (2012), CREs ask students to revisit a topic they have written on and see it through the eyes of another student. This indirectly provides them with an opportunity for self-examination. One might expect the Interpretation, Objections, and Evidence subscores on subsequent essays to benefit from the exercise.

Like CREs, CTEs are also domain specific. CTEs used in this study are a series of scenarios depicting moral and political decision-making that the students were asked to logically analyze. Students were then asked (as homework) to identify the moral principles in play and think about the implications of their opinions. Other types of exercises included the analysis of hidden premises in pared down arguments found in internet memes. The students were asked to identify flaws in the logic of the argument proposed, imagine what would have to be true for the argument to be sound, reconstruct the argument in a more plausible way, and assess the implications. One would expect CTEs to have an impact on all the dimensions of critical thinking.

Course Designs

Aside from comparing the relative impact of rival strategies for promoting CT in the course, the paper also has the opportunity to examine the impact of different course designs.

One plausible route to improving CT is to try to increase the overall quantity of CT assignments. If proficiency is a function of practice, then more practice is likely better. In political science, this "scaffolding" approach is endorsed (in various forms) by Fitzgerald and Baird (2011), Ewell and Rodgers (2014) and Mulcare and Schwedel (2015). On the other hand, if students only have so many hours to devote to a course, it is possible that a surfeit of lower stakes assignments causes less careful and well thought out reasoning for each assignment – and especially for the CT essays.

Another feature of the course design instructors have control over is the weight of each assignment in the final grade. The literature on test taking and student effort suggests that the stakes of the test impact scores. If a small number of points are in play, many students reduce their effort and test and writing assignments scores fall. (Barry et. al. 2010, Wise and Demars 2005, and Elbow 1997). Sewell (2004) also found that dropping a quiz or homework grade negatively affected performance on a final exam. But CT scores are often thought of as measures of ability, not effort, so a question arises as to whether CT scores are sensitive to the stakes in play. Both the percentage value of an essay in the final grade and whether or not it was announced that the worst essay would be dropped are tested.

Other course design features tested include: the presence or absence of exams and whether or not course slides were made available to students.

Key Control Variables

A group of control variables was selected from several meta-analyses of CT studies in higher education (Abrami. et. al 2015, Pithers and Soden 2000, Tiruneh, Verbugh, and Elen 2014) as well as from plausible theoretical assumptions about student behavior. These variables seek to account for three principal sources of variation in CT scores: the individual characteristics of the student, the specific characteristics or context of the assignment, and features of the course-design itself – the latter two being the most pedagogically interesting since they can be manipulated by the instructor.

Control variables collected include: GPA, ACT scores, the percentage of online readings accessed, the percentage of online slides accessed, attendance grades, gender, age, the number of prior essays, and the number of pages of reading since the last essay. Cross-correlations for all these variables can be found in Table 1 and Table 2. Descriptive statistics, collection methods, and a brief discussion of validity and reliability issues pertaining to each variable are discussed in Appendix C.

Finally, when levels of CT are being considered, it is important to adjust scores for topic difficulty. Each of the 23 topics given was therefore assigned a dummy variable so as to produce topic insensitive estimates.

Methods

If CT is subject-specific, then the best predictors in the literature from other disciplines (or even other subfields of political science) might not all apply for political theory. A backwards elimination method and a (forward) stepwise regression method were therefore used to create regression specifications that combine the best predictors available from the pool of independent variables. Because of the similarity of the results, only the backwards elimination method is presented[4].

The only theoretical stricture imposed on the stepwise procedure was the inclusion of the bloc of topic dummy variables.

Since the goal of the study is not to obtain the most parsimonious model in order to leverage a small n relative to a large number of potential regressors, the stepwise selection rules adopted were relatively inclusive. Each model was constructed using a p value cut off of .10[5].

[4] The forward stepwise specifications are available upon request. The differences are minor.
[5] Other decision criteria were considered, including minimum AICc and BIC, but were rejected on the grounds that parsimony and fit were less important than determining whether there was some statistically significant that relationship could be detected for the purposes of future investigation.

Results

Preparatory Assignments, Volume of Assignments, and Manipulating the Stakes

The view that the stakes matter for CT scores fares well in this investigation. The aggregate measure (Table 5) and all the subcategory scores bar one are negatively associated with the stakes of the assignment (Table 3; Table 4). Announcing that the worst essay grade from the semester will be dropped is associated with a further decline in CT scores on the Interpretation category.

The number of CT assignments in a semester is also negatively associated with the aggregate measure of CT (Table 5) as are the measures for selecting evidence and assessing implications (Table 3). In the latter two cases, effect sizes are important, with the number of assignments accounting for a greater quantity of variation than even GPA or ACT scores.

Though more assignments overall are apparently not a path to better scores, subject specific CT assignments fare better than generic ones. CREs help some subcategories (Objections, Evidence, Conclusions) as do CTEs (Conclusions, Objections). The Prior Essays variable was also positive and significant for the Focus score. LGEs, on the other hand, fare poorly: always carrying a negative sign and sometimes appearing to be significantly associated with lower CT scores (Evidence, Conclusion).

Essay Topics and Prompts

If we focus on the subcategories, we can see that topics account for between 8.3 and 15.7% of the variation in student CT scores, by far the largest single source of variation[6] (Table 3; Table 4). The magnitude of the effects for individual topics was not usually trivial, routinely pushing scores 20 or 30 points or more in either direction, though the sample sizes for each topic were sometimes small.

Other results

As expected, a student's GPA at the start of the semester, ACT Scores, and the percentage of readings accessed all have a strong association with CT scores overall and for each dimension (Table 3; Table 4; Table 5). The sign for the dummy variable indicating female gender is also negatively associated with CT scores overall and the Objections score. On the other hand, class attendance, the number of pages assigned between assignments, hours of college credit, and majoring in political science are never statistically significant.

That being a political science major and hours of prior college credit are not significantly correlated with CT scores in political theory probably boosts the case for the existence of a domain-specific moral reasoning skill distinct from other forms of reasoning practiced in political science classes or elsewhere in the college curriculum.

[6] These figures are an average for all topics. Some individual topics may account for more, others less – hence the need for more fine grained analysis than I can give the topic here.

Discussion

A Theory of Rational Irrationality

One of the most salient results of the study is that students appear to adjust their level of CT to the stakes of each assignment. That students reduce their level of effort as the stakes in play decline is not a result that will surprise most instructors. What is perhaps not as obvious is that CT scores also drop as the stakes are lowered. But of course thinking clearly requires effort and discipline that is often in scarce supply.

The idea that individuals vary the quality of their reasoning based on the costs and benefits of thinking well is an idea pioneered by economist Bryan Caplan. Caplan calls this phenomenon "rational irrationality," treating "cognitive inadequacy as a choice" (2007, 123) at least within some bounded region. Caplan draws a relationship between two familiar categories from epistemology: "epistemic rationality" and "instrumental rationality."

Huemer (2016) defines epistemic rationality as "forming beliefs in truth-conducive ways – accepting beliefs that are well-supported by evidence, avoiding logical fallacies, avoiding contradictions, revising one's beliefs in light of new evidence against them and so on," noting in passing that "This is the kind of rationality that books on logic and critical thinking aim to instill." (461) Instrumental rationality, on the other hand, is defined as "adopting suitable means to one's ends" (Stanford Encyclopedia of

Philosophy) or "judging an action by its anticipated consequences" (Hoppe 2003: 120), Because exercising epistemic rationality is costly in terms of time and effort, Caplan thinks it will not always be instrumentally rational to do so.

This trade-off between instrumental and epistemic rationality may be in evidence when students approach CT assignments. When the price of not thinking clearly is low, students respond by indulging in less rigorous thinking - for example by applying little scrutiny to facts, launching into chains of reasoning that flatter their self-image, or failing to examine alternative perspectives to their own. When the price of intellectual indulgence is higher, they respond by tightening up their standards of evidence and crafting own arguments more carefully.

If this theory of "rational irrationality" is correct, then important implications for course design and program assessment follow.

Low-Stakes Preparation, Assignment Overload, and Course Design

As discussed in the literature review, a popular educational strategy consists of assigning low-stakes critical thinking practice assignments to build up student CT capabilities. If students are instrumentally rational about deploying their CT skills, this strategy could backfire in at least three ways.

One possibility is that small stakes practice means low quality practice. As with other skill-based activity, poor quality practice can lead to poor quality performance. Since the

total grade in the class can only ever add up to 100%, this would help make sense of the decline in CT scores as the number of assignments goes up. To practice well, students would need to be properly incentivized, and this would put a cap on how many scaffolding exercises an instructor could assign without seeing the quality of CT begin to drop (through declining stakes).

A second possibility is less dramatic, and makes better sense of the results pointing to CTEs and CREs having some positive effect on CT scores. On this view, low stakes practice could be helpful, but only if it leaves enough weight to the assignments where CT is measured. In some course designs, the quantity of CT exercises could simply be crowding out the essays, leading to lower measured CT scores. CT skills could be improving at the same time as the disposition to showcase those skills is being depressed by the lower stakes attached to the essays.

Finally, thinking well may also simply take time. Too many assignments and students might not have enough time between assignments to devote to thinking about their essays. In that case, the proper spacing of assignments and especially giving students enough time to think about the essay topic may help raise CT scores.

Clearly more research is needed to disentangle these different hypotheses, but whatever turns out to be the case, more low-stakes practice is not always pedagogically optimal.

Manipulating the Stakes, Essay Topics, and Program Assessment

Variations in stakes and topics should also give us pause when thinking about program assessment. If Department A conducts an exit exam with a CT component using a standardized test worth some percentage of the final grade in a senior seminar and Department B requires the very same exit exam to graduate, but fails to attach a consequence to low performance, statistics comparing the programs will fail to accurately measure the differences in their graduates.

The same will be true if students are given a CT exam upon entering college with one set of stakes attached, but then given a CT exit exam in their senior year with different stakes attached. This second scenario is further complicated by the fact that, controlling for individual and course characteristics, different topics produce very different average scores, as well different degrees of subscore variation. One can readily imagine a student having made great strides in their ability to articulate alternative perspectives, yet have that improvement go unmeasured if the post-test is much more difficult on that particular dimension of critical thinking.

Does it matter what the stakes are so long as they are consistent? According to the theory of rational irrationality, the answer is yes. If the stakes are low, then only students who have strong internal motivation will exercise the intellectual discipline necessary to achieve high scores. If the stakes are high, both the internally motivated *and* those capable of being externally motivated towards higher levels of CT will achieve higher scores. If this second population is non-trivial in size, instructors and departments may be underestimating their students' capabilities (and even their own performance) by some

margin when they attach no real stakes to CT assessments. The coefficients obtained in this study suggest that the effect of changing the stakes is non-trivial. Doubling the stakes of an essay from 5 to 10% of the final grade in the class could yield up to 20 extra percentage points in measured CT.

Several researchers have suggested a need to focus higher education on developing students who have the disposition (as well as the ability) to engage in critical thought (Giancarlo and Facione 2001; Pascarella and Terenzini 2005). Looking at the findings from this study, there could indeed be large gains in measured CT from improving the disposition to exercise CT at low stakes.

Directions for Further Research

Aside from experimental confirmation of the impact of stakes on measured CT scores, several other areas for further research stand out.

Differences between high- and low- achieving students are of special interest. This study measures the average effect of different factors across all students, but there may be subgroups of students whose demonstrated CT is particularly helped or harmed by different contexts and instructional strategies. Other studies have found this sort of effect among psychology students. (Williams et. al 2003, 2004)

The negative coefficients for female students also merits further research. In other contexts and disciplines, gender has been found to have both positive, negative, and insignificant associations with CT. (Facione 1990, Walsh and Hardy 1999, Rudd Baker and Hoover 2000, Friedel et. al 2008) It would be helpful to understand the mechanism through which gender influences CT scores in political theory.

One hypothesis (supported by informal class surveys of interest and the fact that average scores by topic vary sharply according to gender) is that the curriculum in the class is simply (on average) less interesting to female students. This would support Nussbaum's concern that political theory has done a poor job of catering to the interests of women (1999). Another explanation might be that the types of moral reasoning that are normal and accepted in political theory are excessively "masculine." (Held 1987) Either of these criticisms alone, if substantiated, would be motive enough for rethinking the traditional curriculum in this and other theory courses.

Finally, Fitzgerald and Baird (2011) suggest that topics for CT assignments on normative issues should fall outside domains that are likely to be familiar and politically charged for them. By assigning topics that are highly ideologically charged (where students may be tempted to "consume" ideas or beliefs that they prefer to believe), one would expect CT scores to fall. Examining what makes some topic averages worse than others more generally seems like a promising avenue for future research and may provide some tangible payoffs to researchers and programs looking to develop topic-insensitive measures of CT for assessment purposes.

Works Cited

Bailin, Sharon. 2002. "Critical Thinking and Science Education." *Science & Education* 11(4): 361–75.

Barry, Carol L. et al. 2010. "Do Examinees Have Similar Test-Taking Effort? A High-Stakes Question for Low-Stakes Testing." *International Journal of Testing* 10(4): 342–63.

Bataineh, Ruba Fahmi, and Lamma Hmoud Zghoul. 2006. "Jordanian TEFL Graduate Students' Use of Critical Thinking Skills (as Measured by the Cornell Critical Thinking Test, Level Z)." *International Journal of Bilingual Education and Bilingualism* 9(1): 33–50.

Blings, Steffen, and Sarah Maxey. 2016. "Teaching Students to Engage with Evidence: An Evaluation of Structured Writing and Classroom Discussion Strategies." *Journal of Political Science Education* 13(1): 15–32.

Bok, Derek. 2008. *Our Underachieving Colleges: A Candid Look at How Much Students Learn and Why They Should Be Learning More*. Princeton: Princeton University Press.

Brown, J. D., T. Hilgers, and J. Marsella. 1991. "Essay Prompts and Topics: Minimizing the Effect of Mean Differences." *Written Communication* 8(4): 533–56.

Caplan, Bryan Douglas. 2008. *The Myth of the Rational Voter: Why Democracies Choose Bad Policies*. Princeton, NJ: Princeton Univ. Press.

Carbonaro, William. 2005. "Tracking, Students' Effort, and Academic Achievement." *Sociology of Education* 78(1): 27–49.

Cavdar, Gamze, and Sue Doe. 2012. "Learning through Writing: Teaching Critical Thinking Skills in Writing Assignments." *PS: Political Science & Politics* 45(02): 298–306.

Condon, William, and Diane Kelly-Riley. 2004. "Assessing and Teaching What We Value: The Relationship between College-Level Writing and Critical Thinking Abilities." *Assessing Writing* 9(1): 56–75.

Elbow, Peter. 1997. "High Stakes and Low Stakes in Assigning and Responding to Writing." *New Directions for Teaching and Learning* 1997(69): 5–13.

Ennis, Robert H. 1989. "Critical Thinking and Subject Specificity: Clarification and Needed Research." *Educational Researcher* 18(3): 4.

Ewell, William Henry, and Robert R. Rodgers. 2014. "Enhancing Student Preparedness for Class through Course Preparation Assignments: Preliminary Evidence from the Classroom." *Journal of Political Science Education* 10(2): 204–21.

Facione, N C, P A Facione, and C A Sanchez. 1994. "Critical Thinking Dispositions as a Measure of Competent Clinical Judgment: the Development of the Critical Thinking Disposition Inventory. ." *Journal of Nursing Education* 33(8): 345–50.

Facione, Peter A. 1990. *The California Critical Thinking Skills Test--College Level. Technical Report No. 3. Gender, Ethnicity, Major, Ct Self-Esteem, and the Cctst*. Education Resources Information Center.

Facione, Peter A., and Noreen C. Facione. 1994. "Critical Thinking Ability: A Measurement Tool." *Assessment Update* 6(6): 12–13.

Fitzgerald, Jennifer, and Vanessa A. Baird. 2011. "Taking a Step Back: Teaching Critical Thinking by Distinguishing Appropriate Types of Evidence." *PS: Political Science & Politics* 44(03): 619–24.

Friedel, Curt et al. 2008. "Overtly Teaching Critical Thinking and Inquiry-Based Learning: A Comparison of Two Undergraduate Biotechnology Classes." *Journal of Agricultural Education* 49(1): 72–84.

Gaff, Jeffrey, and Jack Meacham. 2014. Learning Goals in Mission Statements: Implications for Educational Leadership *Learning Goals in Mission Statements: Implications for Educational Leadership*. Association of American Colleges & Universities. rep. https://www.aacu.org/publications-research/periodicals/learning-goals-mission-statements-implications-educational (July 21, 2017).

Gelder, Tim Van. 2005. "Teaching Critical Thinking: Some Lessons From Cognitive Science." *College Teaching* 53(1): 41–48.

Giancarlo, Carol Ann, and Peter A. Facione. 2001. "A Look Across Four Years at the Disposition Toward Critical Thinking Among Undergraduate Students." *The Journal of General Education* 50(1): 29–55.

Halpern, Diane F. 2001. "Assessing the Effectiveness of Critical Thinking Instruction." *The Journal of General Education* 50(4): 270–86.

Harvey, A. C. 1976. "Estimating Regression Models with Multiplicative Heteroscedasticity." *Econometrica* 44(3): 461–66.

Held, Virginia. 1987. "Non-Contractual Society: a Feminist View." In *Science, Morality, and Feminist Theory*, Canadian Journal of Philosophy, Supplementary Volume, eds. Marsha P. Hanen and Kai Nielsen. Calgary: University of Calgary Press.

Hoppe, Robert. 2003. "Dealing with Multiple Rationalities: A Rejoinder to Ventriss, Tenbensel and Snellen." *Administrative Theory & Praxis* 25(1): 119–27. http://www.jstor.org/stable/10.2307/25610593?ref=search-gateway:e554e4e2a58720e747c1a2eeb18b2d07 (March 21, 2017).

Howard, Larry W., Thomas Li-Ping Tang, and M. Jill Austin. 2014. "Teaching Critical Thinking Skills: Ability, Motivation, Intervention, and the Pygmalion Effect." *Journal of Business Ethics* 128(1): 133–47.

Howard-Snyder, Frances, Daniel Howard-Snyder, and Ryan Wasserman. 2012. *The Power of Logic*. New York: McGraw-Hill.

Huemer, Michael. 2016. "Why People Are Irrational About Politics." In *Philosophy, Politics, and Economics: an Anthology*, eds. Jonathan Anomaly, Geoffrey Brennan, Michael C. Munger, and Geoffrey Sayre-McCord. New York: Oxford University Press. essay, 456–67.

"Job Outlook 2016: The Attributes Employers Want to See on New College Graduates' Resumes." http://www.naceweb.org/s11182015/employers-look-for-in-new-hires.aspx (March 19, 2017).

Kelly-Riley, Diane, Gary Brown, Bill Condon, and Richard Law. 2008. Washington Center University Critical Thinking Project: Resource Guide *Washington Center University Critical Thinking Project: Resource Guide*. Washington State University Critical Thinking Project. rep.

Klein, S., R. Benjamin, R. Shavelson, and R. Bolus. 2007. "The Collegiate Learning Assessment: Facts and Fantasies." *Evaluation Review* 31(5): 415–39.

Marks, Michael P. 2008. "Fostering Scholarly Discussion and Critical Thinking in the Political Science Classroom." *Journal of Political Science Education* 4(2): 205–24.

Mason, Mark. 2007. "Critical Thinking and Learning." *Educational Philosophy and Theory* 39(4): 339–49.

McPeck, John E. 1981. *Critical Thinking and Education*. New York: St. Martin's Press.

Moore, Matthew J. "How (and What) Political Theorists Teach: Results of a National Survey." *Journal of Political Science Education* 7, no. 1 (2011): 95-128. doi:10.1080/15512169.2011.539921.

Mulcare, Daniel M., and Allan Schwedel. 2016. "Transforming Bloom's Taxonomy into Classroom Practice: A Practical Yet Comprehensive Approach to Promote Critical Reading and Student Participation." *Journal of Political Science Education* 13(2): 121–37.

Musekamp, Frank, and Jacob Pearce. 2016. "Student Motivation in Low-Stakes Assessment Contexts: an Exploratory Analysis in Engineering Mechanics." *Assessment & Evaluation in Higher Education* 41(5): 750–69.

Nussbaum, Martha Craven. 1999. *Sex & Social Justice*. Oxford: Oxford University Press.

Olsen, Jonathan, and Anne Statham. 2005. "Critical Thinking in Political Science: Evidence from the Introductory Comparative Politics Course." *Journal of Political Science Education* 1(3): 323–44.

Omelicheva, Mariya Y. 2007. "Resolved: Academic Debate Should Be a Part of Political Science Curricula." *Journal of Political Science Education* 3(2): 161–75.

Oros, Andrew L. 2007. "Let's Debate: Active Learning Encourages Student Participation and Critical Thinking." *Journal of Poltiical Science Education* 3(3): 293–311.

Pascarella, Ernest T., and Patrick T. Terenzini. 2005. *How College Affects Students*. San Francisco, CA: Jossey-Bass.

Paul, R. 1991. "Critical Thinking: What Every Person Needs To Survive in a Changing World." *NASSP Bulletin* 75(533): 120–22.

Phillips, John L., and Laura Mcmillian. 2010. "Liberal Civic Education, Religious Commitment, and the Spillover Thesis: What Psychology Can Teach Us." *Politics and Religion* 4(01): 27–48.

Pithers, R.t., and Rebecca Soden. 2000. "Critical Thinking in Education: a Review." *Educational Research* 42(3): 237–49.

Resnick, Daniel, and Natalie Peterson. 1991. "Evaluating Progress toward Goal Five: A Report to the National Center for Education Statistics." https://eric.ed.gov/?id=ED340764.

Rudd, Rick, Matt Baker, and Tracy Hoover. 2008. "Undergraduate Agriculture Student Learning Styles and Critical Thinking Abilities: Is There a Relationship?" *Journal of Agricultural Education* 41(3): 2–12.

Sewell, Ellen. 2004. "Grade Dropping: An Empirical Analysis." *The Journal of Economic Education* 35(1): 24–34.

Smith, Gerald. 2002. "Are There Domain-Specific Thinking Skills?" *Journal of Philosophy of Education* 36(2): 207–27.

Stein, Barry et al. 2010. "Faculty Driven Assessment of Critical Thinking: National Dissemination of the CAT Instrument." *Technological Developments in Networking, Education and Automation*: 55–58.

Terenzini, Patrick T., Leonard Springer, Ernest T. Pascarella, and Amaury Nora. 1995. "Influences Affecting the Development of Students' Critical Thinking Skills." *Research in Higher Education* 36(1): 23–39.

Tiruneh, Dawit T., An Verburgh, and Jan Elen. 2014. "Effectiveness of Critical Thinking Instruction in Higher Education: A Systematic Review of Intervention Studies." *Higher Education Studies* 4(1).

Walsh, Catherine M, and Robert C Hardy . 1999. "Dispositional Differences in Critical Thinking Related to Gender and Academic Major." *Journal of Nursing Education* 38(4): `49–155. http://eric.ed.gov/?id=EJ583065 (April 13, 2017).

Williams, Leonard, and Mary Lahman. 2011. "Online Discussion, Student Engagement, and Critical Thinking." *Journal of Political Science Education* 7(2): 143–62.

Williams, Robert L., and Stephen L. Worth. 2001. "The Relationship of Critical Thinking to Success in College." *Inquiry: Critical Thinking Across the Disciplines* 21(1): 5–16.

Williams, Robert L., Renee Oliver, and Susan L. Stockdale. 2004. "Psychological Versus Generic Critical Thinking as Predictors and Outcome Measures in a Large Undergraduate Human Development Course." *The Journal of General Education* 53(1): 37–58.

Wise, Steven L., and Christine E. Demars. 2005. "Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions." *Educational Assessment* 10(1): 1–17.

Appendix A: The Grading Rubric

The papers in the class were graded along 6 dimensions of critical thinking, inspired by the grading rubric developed by the Washington State University Critical Thinking Project:

[Insert Table A1]

Each of these 6 categories is scored on a spectrum from 0 to 6 where 0 means absent, 1 means minimal, 2 means "emerging," 3 means "developing," 4 means "competent," 5 means "effective," and 6 means "mastering."

Although the grading rubric used in this class departs from the original WSU categories, the WSU critical thinking project allows and even encourages variation and adaptation of the assessment instrument to different subjects. (Kelly-Riley. et. al 2008)

The WSU rubric separates Issue Identification from Identifying Contexts and Assumptions, whereas I combine them. The WSU Rubric combines Interpretation and Evidence, whereas I separate them. The justification for doing this initially was to make room for textual interpretation – which is a domain much cherished by political theorists. Combining Issue Identification and Identifying Context and Assumption was done because in normative contexts it is often hard to separate "determining what the issue is" from "determining what the different positions on the issue are."

This rubric also contains no Communication category. This category was scored but is not reported here. The reasons for this are twofold. On theoretical grounds it is not clear that command of the English language is conceptually related to CT. Moreover, a factor analysis conducted in earlier versions of the project determined that while the sources of variation in the six CT categories overlapped quite a bit – a single distinct source of variation could account for 94% of the variation in Communication scores. The distinction between CT and writing skill is something the authors of the WSU rubric also seem to concede (Condon and Kelly-Riley, 2004)

For ease of interpretation, scores on the 7-point scale (0-6) are converted to a 100 point scale in the study.

Appendix B: Test of Intercoder Reliability

Methodology

A post-test study of intercoder reliability was conducted between the original grader and two senior undergraduates applying to graduate school in political science. Although ideally a fully random and sample of papers would be used, time and budget constraints led to choosing a random sample of 4 topics – each of which had been used over multiple semesters. The topics used were Topics 103, 107, 110, and 112. It was felt that using the full range of topics would make the burden of grading unduly onerous on the undergraduates. Instead, three random semesters were selected and two anonymous random papers from each semester

were selected using an online random number generator. The test therefore comprised a total of 24 papers per grader (6 per topic).

In this instance, the principal worry is about overall agreement between different coders (especially that high rating and low ratings were consistent). Two measures are created. Mean Absolute Differences (MAD) and Pearson's r. The first establishes how far away, on average, each rating is from another. The second is a measure of correlation for each of the dimensions of critical thinking (as well as overall). Here obviously the higher the correlation, the better. [Table B1]

Discussion

The overall critical thinking score is satisfactory (considering that two of the graders were not subject matter experts) with a MAD score of 4.38 out of a possible 43 points (10.42% deviation) and an overall Pearson's r of .61. The thesis category appears to be the most problematic. For that category, Pearson's r correlations are weak and Mean MAD for the graders are over 20%. The evidence category has weak correlations but much better MAD scores, indicating close scores but uneven co-variation between the graders – which is somewhat puzzling. All other categories exhibit statistically significant covariation and acceptable MAD scores.

Appendix C – The Independent Variables

[Table C1]

GPA: Studies have found that students with higher GPAs do better on critical thinking measures. (Bataineh and Zghoul 2006; Howard, Li-Ping Tang, and Austin 2014) The inclusion of GPA is also important for another reason: weaker students drop out of the class as the semester progresses, leading to fewer observations for those students than for stronger students. Controlling for GPA will limit the effect of any bias arising out of missing data.

ACTs: Prior studies have found a relationship between ACT scores and critical thinking over and above the impact of GPA. (Howard, Li-Ping Tang, and Austin 2014) A downside of using ACT scores as a control variable in this particular study is that many students at the university are not required to take any kind of standardized test for admissions. Using ACT scores causes 146 observations to drop from the study (35% of the total). The populations are similar in most respects. However, two statistically significant differences are worth noting: the mean age of excluded population is much higher (20.83 vs. 27.75) and the gender distribution skews more female (60.59% male vs. 51.41% male).

In the end, the decision was taken to include ACT scores in the study on the grounds that even if the results skewed the population of students towards more traditional students, this might make the study more comparable to other studies of critical thinking at the college level. A limitation of the study is that the critical thinking habits of older non-traditional students must be set aside for further study and that generalization of any

insights to this population would be hazardous at best. Three ACT scores were collected (English, Math, and Reading). These were averaged into a composite score.

Reading Percentage/Slide Percentage: The reading percentage and slide percentage variables measure the percentage of online readings and slides accessed from the beginning of class up until a given scored essay[1]. Although we cannot be sure that a student who accessed the readings/slides actually read them, we can be fairly confident that a file that was never visited was never read.

Only a portion of the course readings are online. The reading measure does not capture engagement with the textbooks and only very few classic texts. Some students may read regularly from the textbooks but access online content only seldom. Other students may never buy the textbooks, but reliably access any content available for free online. All students who did not have access to the slides during the semester in which they took the course were coded as zero.

Age: Age proxies for maturity and experience. Age in the study is measured in years. If a student turned a different age during the semester, they were coded as having the age they began the course with.

[1] A related measure, the percentage of readings accessed since the last assignment was also computed, but dropped from the study since it produced very similar overall results

Attendance: Attendance grades are a combination of class attendance and attendance at extra-curricular events. Every class missed equals 5 points less on the attendance grade. The failure of class attendance to strongly correlate with critical thinking measures in the study could be due to its imprecise measurement.

Gender: Students who requested to be treated as a belonging to a different gender than the one suggested by their first name were assigned the gender of their choice in the study (3 of 151), but this had no impact on results.

Hours of College Credit: An alternate measure of college experience was also constructed in the form of an ordinal variable separating students into First Year Students, Sophomores, Juniors, Seniors, and Super-seniors (over 120 college credits) but did not prove any more helpful in predicting CT.

Exams: Exams were short answer exams, covering terms and concepts from the readings and lectures, as well as some questions asking students to apply principles to novel situations.

Previous Essays/Critiques/CTEs: These variables track the student, not the number of assignments to date in that semester. For example, if a student fails to hand in essay #1, the variable is coded 0 for that student even though all other students might be coded 1 for that essay. CTEs were worth between 2.5 and 5% of the final grade. Critiques were worth 5-10% of the final grade.

Appendix D: Discussion of Limitations Due to Cross-Sectional Design

One danger presented by the multiple regression methodology for data spanning across semesters in this way is that individual measures of critical thinking are not independent of one another. Although some effort was made to control for class and assignment level variables, it is possible that each semester exerted an influence not captured by the control variables. If the main sources of semester variation are captured this should not be a problem – but omitted variable bias is always a threat[2].

A Durbin Watson test was effected for each model to check for serial autocorrelation, with results between 2.06 and 2.28 showing no statistically significant autocorrelation (Table D1). This suggests that the problems associated with non-independence of observation are minimal.

Visually inspecting the pattern of residuals through a Q-Q plot reveals well-behaved residuals for the Focus, Interpretation, Objections, Evidence, and Overall Critical Thinking. Those for the Thesis and Conclusion models present potential problems for the assumption that errors are normally distributed. A Shapiro-Wilk test confirms this. (Table D1) This could suggest bias in the coefficients for those models due to one or more omitted variables. It could also signify non-random sources of bias in grading.

[2] A regression using a semester dummy variable and a standard set of individual and assignment level controls finds that semester level effects are significant in the aggregate and account for around 3.5% of the variation in aggregate critical thinking scores.

The problematic models were re-run using log-linear variance regression (Harvey 1976) to simultaneously fit the expected value of the dependent variables and studentize the residuals by fitting the variance of the responses (thus producing a model with heteroscedasticity-consistent standard errors). The percentage of slides variable drops from the Conclusion model. The essay value variable drops from significance in the Thesis model. Otherwise, the results are consistent.

Table 1
Correlation Matrix – Part 1 (Pearson's r)

	GPA	ACT Avg	Age	Gend.	Reading%	Slide %	Attend	Hours	Maj.
GPA	1.00								
ACT Avg	0.39***	1.00							
Age	-0.24***	-0.12**	1.00						
Gender	0.08	0.03	-0.13**	1.00					
Reading %	0.28***	-0.03	0.06	0.18***	1.00				
Slide %	-0.05	-0.17***	0.19***	0.00	0.25***	1.00			
Attend.	0.31***	0.00	0.05	0.02	0.28***	0.15***	1.00		
Hours	0.15**	0.07	0.44	0.00	0.09	0.10**	0.08	1.00	
Major	0.11*	-0.03	0.09	-0.02	0.11*	0.06	0.09	0.20**	1.00
Essay Drop	-0.10*	-0.16***	-0.06	0.11*	0.20***	0.49***	0.04	0.03	0.09
Essay value	-0.03	0.11*	0.15**	-0.09	-0.12**	-0.08	0.17***	-0.10*	0.13**
# of Assignments	-0.10*	-0.01	-0.07	0.06	0.13**	0.48***	-0.06	0.08	0.15***
LGs	0.03	-0.03	0.13**	-0.14**	-0.16***	-0.14**	-0.06	-0.01	-0.16***
Slide dummy	-0.10*	-0.06	0.03	-0.01	0.23***	0.61***	0.17***	-0.06	0.04
# of Pages	0.04	0.02	-0.03	0.05	0.09	0.02	0.07	-0.03	0.04
Exams	-0.05	-0.09	0.09	0.02	0.08	0.40***	0.00	0.06	-0.05
CTEs	-0.02	-0.10	0.08	0.04	0.17***	0.54***	0.00	0.13**	-0.02
Prev. Essays	0.00	0.00	-0.01	-0.02	-0.08	0.08	-0.01	0.07	0.00
CREs.	0.08	0.11	-0.12**	-0.12**	-0.17***	-0.37	0.01	-0.03	0.08

p<.1 **: p<.05 ***p<.01

Table 2
Correlation Matrix – Part 2 (Pearson's r)

	Essay value	# of Assignments	LGs	Slide dummy	# of Pages	Exam	CTEs	Prev. Essay	CREs
Essay Drop	-0.33***	0.50***	-0.48***	0.55***	0.17***	0.31***	0.38***	0.09	-0.28***
Essay value	1.00	0.07	-0.21***	0.18***	0.23***	-0.20***	-0.20***	-0.12**	0.11*
# of Assignments		1.00	-0.59***	0.56***	0.13**	0.22***	0.66***	0.20***	-0.25***
LGs			1.00	-0.38***	-0.29***	0.25***	0.00	-0.12**	-0.04
Slide dummy				1.00	0.30***	0.34***	0.40***	-0.01	-0.23***
# of Pages					1.00	0.08	-0.06	-0.13**	0.12**
Exams						1.00	0.69***	0.29***	-0.33***
CTEs							1.00	0.34***	-0.39***
Prev. Essays								1.00	0.47***
CREs.									1.00

p<.1 **: p<.05 ***p<.01

Table 3
Stepwise Backward Elimination Regression Results: Evidence, Conclusion, Objections

	Evidence			Conclusion			Objections		
	B	*SE*	(η^2/r^2)	*B*	*SE*	(η^2/r^2)	*B*	*SE*	(η^2/r^2)
Intercept	-1.72	16.87		10.20	18.78		-69.04***	17.55	
TopicID (Prob>F)	.100*		0.085	.137		0.083	<.001***		0.157
ACTs	1.20***	0.29	0.044	1.05***	0.33	0.027	0.77**	0.36	0.011
GPA	6.27***	2.14	0.022	4.12*	2.40	0.008	15.98***	2.53	0.097
Essay value	4.90***	1.68	0.022	5.47***	1.77	0.026	3.59*	1.84	0.009
Reading %	0.13***	0.05	0.020	0.15***	0.05	0.022			
CREs	5.22**	2.48	0.012	7.02***	2.38	0.023	6.38**	2.61	0.014
Slide dummy[1]	7.97**	3.68	0.012						
Exams	18.42**	7.38	0.016						
# of Assignments	-6.20***	1.43	0.049	-5.83***	1.70	0.032			
LG Dummy[1]	-7.21*	4.35	0.007	-10.81**	4.38	0.016			
CTEs				13.82***	3.73	0.037	6.64**	2.95	0.012
Slide %				12.65**	5.79	0.013			
Gender [F]							-4.11***	1.47	0.019
Age									
Prev. Essays									
Essay Drop									
# of Pages									
Attendance									
Hours									
Major									
Adj. R^2		.256			.232			.305	
R^2		.340			.319			.375	
N		286			286			289	

***: p<.1; **p<.05; ***:p<.01**

Table 4
Stepwise Backward Elimination Regression Results: Thesis, Focus, Interpretation

	Thesis			Focus			Interpretation		
	B	*SE*	*(η^2/r^2)*	*B*	*SE*	*(η^2/r^2)*	*B*	*SE*	*(η^2/r^2)*
Intercept	-10.07	16.75		-52.25***	19.42		-19.63	19.12	
TopicID (Prob>F)	.012**		0.128	.010***		0.127	.100*		0.092
ACTs	0.97***	0.35	0.022	0.89**	0.35	0.019			
GPA	10.14***	2.50	0.048	7.87***	2.67	0.025	11.84***	2.33	0.069
Essay value	3.08*	1.76	0.009				4.19**	1.95	0.012
Reading %				0.14***	0.05	0.020	0.20***	0.05	0.040
CREs									
Slide dummy[1]				-7.07**	3.26	0.013			
Exams							10.72**	4.97	0.012
# of Assignments									
LG Dummy[1]									
CTEs									
Slide %									
Gender [F]									
Age				2.30***	0.76	0.026			
Prev.Essays				6.81**	3.16	0.013			
Essay Drop							-5.58**	2.79	0.011
# of Pages									
Attendance									
Hours									
Major									
Adj. R^2	.151			.183			.232		
R^2	.229			.268			.310		
N	292			289			289		

***: p<.1; **p<.05; ***:p<.01**

Table 5
Stepwise Backward Elimination Regression Results: Aggregate Critical Thinking

	All Regressors			Selected Model		
	B	SE	(η^2/r^2)	B	SE	(η^2/r^2)
Intercept	-35.56*	20.18		-19.71	12.48	
TopicID (Prob>F)	.682			.453		0.056
ACTs	0.82***	0.24	0.028	0.84***	0.22	0.075
GPA	9.67***	1.97	0.056	9.07***	1.62	0.045
Essay value	4.91***	1.57	0.023	4.16***	1.10	0.035
Reading %	0.13***	0.04	0.029	0.15***	0.03	0.016
Gender[F]	-2.35**	0.92	0.015	-2.31**	0.89	0.011
# of Assignments	-4.17***	1.46	0.019	-1.55**	0.73	0.035
CREs	3.49	2.23	0.006			
Slide dummy[1]	4.22	4.14	0.002			
LG Dummy[1]	-10.11*	5.29	0.009			
Hours	-0.02	0.04	0.000			
Age	1.15*	0.69	0.007			
Prev. Essays	0.45	3.32	0.000			
Dropped Essay[1]	-3.68	4.06	0.002			
Exams	10.19	6.65	0.005			
Attendance	0.02	0.04	0.001			
CTEs	5.08*	2.85	0.007			
Total Slide %	5.31	4.14	0.004			
Pages Since Last Essay	-0.03	0.04	0.001			
Major[1]	-1.56	1.21	0.004			
Adj. R^2		.337			.311	
R^2		.434			.381	
N		286			288	

***: p<.1; **p<.05; ***:p<.01**

Table A.1
Dimensions of Critical Thinking, Aggregate Critical Thinking and Descriptive Statistics

Dimension	Description	Mean	Std. Dev.	N
Issue Identification and Focus	How well did the student show that they understood the question that was being asked? Did the student adequately examine all the concepts put into play by the prompt? Did the student stay on task and answer the question asked?	57.74	23.85	447
Interpretation	Did the student fairly represent different possible positions on the issue to be found in the literature? Did the student more generally draw plausible inferences from the sources they used?	63.60	24.57	446
Evidence	Did the student produce appropriate evidence for claims made in the essay? Did the argument build from plausible assumptions/broadly shared intuitions? Did the student show an awareness of the reliability of different sources for empirical evidence (if such were used)?	53.80	21.04	447
Objections	Did the student consider prominent objections to their thesis or to their assumptions? Were such objections plausibly dealt with?	35.78	26.37	447
Thesis	Does the paper have a clear thesis? Is the position the student's (or is it a simple restatement of a position found in the literature)? Is the argument developed internally consistent – given its premises and assumptions?	66.76	24.59	447
Conclusion	Is the conclusion consistent with the argument presented in the paper? Has the student identified implications for other areas of inquiry or public policy?	73.49	22.69	447
Aggregate CT	An additive measure combining 6 dimensions of critical thinking (each being weighted equally).	58.39	16.58	446

Table B1
Measures of Intercoder Reliability: Mean Absolute Deviations (MAD) and Correlation Coefficients (Pearson's r)

	Undergraduate Grader #1		Undergraduate Grader #2		Faculty Regrade		Mean Overall	
	MAD	Pearson's r	MAD	Pearson's r	MAD	Pearson's r	MAD	Pearson's r
Focus	1.27 (18.22%)	0.55***	1.27 (18.10%)	0.35	1.09 (15.62%)	0.60***	1.21 (17.31%)	0.50***
Interpretation	0.97 (13.89%)	0.59***	1.27 (18.10%)	0.57***	1.30 (18.59%)	0.55***	1.06 (15.22%)	0.57***
Evidence	0.78 (11.22%)	0.20	0.92 (13.17%)	0.25	0.89 (12.78%)	0.30	0.87 (12.41%)	0.25
Objections	1.47 (21.07%)	0.42**	1.20 (17.15%)	0.65***	1.07 (15.31%)	0.70***	1.25 (17.84%)	0.59***
Thesis	1.33 (19.02%)	0.29	1.63 (23.32%)	0.15	1.41 (20.11%)	0.30	1.46 (20.82%)	0.25
Conclusion	1.22 (17.51%)	0.39*	0.96 (13.73%)	0.53***	0.50 (7.09%)	0.81***	0.89 (12.78%)	0.57***
Overall	5.38 (12.81%)	0.56***	4.57 (10.87%)	0.57***	3.18 (7.58%)	0.70***	4.38 (10.42%)	0.61***

***: p<.1; **p<.05; ***:p<.01**

Table C1
Descriptive Statistics: Continuous Variables

	GPA	ACT Avg	Age	Slide %	Reading%	Attend	Hours	Essay value	# of Assignments	# of Pages	Exams	CTEs	CREs
Mean	3.04	22.97	23.26	0.25	74.18	83.99	79.76	8.20	8.24	163.91	0.19	0.51	0.76
Std. Dev	0.64	4.15	5.51	0.34	28.60	22.14	28.80	1.02	2.11	80.51	0.49	1.13	0.94
Min	1	14	18	0	0	0	16	7.14	5	72	0	0	0
Max	4	32	62	1	100	113	190	10	14	477	2	5	3
Collection Method	UR	UR	UR	D2L	D2L	D2L	UR	SYL	SYL & D2L	SYL	SYL	SYL/ D2L	SYL/ D2L
N	443	294	450	445	444	447	446	450	449	447	451	447	451

Descriptive Statistics: Dummy Variables

	Slide Dummy (1=Yes)	Logic Game Dummy (1=Yes)	Major (1=Yes)	Gender (1=Fem)	Dropped Essay (1=Yes)
0	198	239	67	268	175
1	253	208	381	182	275
Collection Method	D2L	SYL	UR	OBS	SYL
N	451	447	448	450	450

UR=University Records; D2L=Online Course Management System; SYL=Syllabus; OBS=Observation

Table D1
Tests of Model Assumptions: Autocorrelation (Durbin Watson), Normal Distribution of Residuals (Shapiro-Wilk), and Associated Probabilities of Accepting Model Assumptions

	Durbin Watson	Prob<DW	Shapiro-Wilk	Prob<W	N
Aggregate CT	2.14	.44	.997	.95	288
Evidence	2.20	.55	.992	.10	286
Conclusion	2.21	.61	.995	.001***	286
Objections	2.06	.19	.995	.041	289
Thesis	2.28	.88	.982	.001***	292
Focus	2.31	.92	.991	.11	289
Interpretation	2.07	.22	.989	.017**	289