Topical Destination: Assessing the Impact of Essay Prompts on Issue Identification and Critical Thinking in Political Theory

Presented at the Annual Meeting of the Northeastern political Science Association
November 9, 2017
Philadelphia, PA

John L. Phillips, Ph.D.
Assistant Professor
Department of Political Science and Public Management
Austin Peay State University
phillipsj@apsu.edu

Abstract: Persuasive essays are still a staple in upper level political science courses and perhaps nowhere more so than in political theory. Intuitively, some topics and prompts are better than others, but are there systematic features of essay topics capable of producing better or worse essays? This study analyzed 800 college essays from a junior level political theory course scored using the WSU Critical Thinking Rubric. Longer prompts predicted better scores across the board as did, to a limited extent, making the topic more personal. Conversely, students struggled more with abstract normative reasoning than conceptual inquiry, especially when the normativity was not explicitly signaled. Open-ended topics and those asking students to process a greater quantity of reading material exerted no discernible effect. The implications for topic choice are discussed.

Introduction

Anyone who has ever had to write a college essay knows that not all topics and prompts are created equal. Some are inspiring, others tedious. Sometimes the type of work that must be produced is very clear. Other times, what the professor expects is a bit of a mystery. In investigations of writing proficiency, topic effects are often discernible (Huot 1990). Both the compatibility of the topic to the students (as measured by past exposure to the subject matter) (Pomplun et. al 1992, Golub-Smith et. al. 1993, Lee and Anderson 2007) and the structural features of the topic such as word length (Hayward 1989), "rhetorical specification of topic purpose and audience" (Oliver 1995), or "two-sidedness" (Hinkel 2002) have been found to exercise an effect. For those of us who teach through writing, but do not teach writing, the quality of the writing is only one of the desired outcomes of an essay assignment, however.

With one or two exceptions (see Fitzgerald and Baird 2011, for example) essay topics have largely escaped the attention of political science and political theory education and quantitative assessment is non-existent. If we survey treatises on teaching in political science, a real dearth of advice on what makes for a good essay topic becomes evident. The *Handbook on Teaching and Learning in Political Science and International Relations* (2015) devotes one chapter out of 37 to writing and says nothing at all about the structural features of a good essay topic. Laura Paquette in *Teaching Political Science to Undergraduates* (2015) says nothing at all about essays or writing either. There are books about writing in political science (for example, Schmidt 2010) but these tend to focus on the aspiring practitioner or the student, not the instructor, and in any case have little reason to discuss what makes a good essay topic for instructional purposes. Finally, there is a little bit of advice on what makes a good dissertation topic (mainly in the form of news articles and blog posts), but the sorts of writing instructors want to elicit at the undergraduate level have little to do with what is required of a graduate dissertation.

Despite all of this, political science majors still aspire to careers where both writing and critical analysis in writing are essential skills. Careful attention to the development of analytical skills through writing is therefore still very much within the purview of teaching in the discipline – though perhaps more so in political theory (see for example Kassiola, 227, 783). Instructors want students to show, among other things, that they master the content of the course, that they can apply theories and concepts discussed in class to new cases, and that they can make a persuasive case in favor of this or that thesis using evidence and logical rigor. It would therefore be to our advantage as instructors to know something about what kinds of essay topics elicit better quality responses from our majors.

This paper explores this question using a data set comprised of 800 blind-graded college essays (collected over a span of 13 semesters) from an introductory political theory class at a midsize public university in the southern United States. The grading rubric for the essay, heavily modeled on the Washington State University critical thinking rubric (Condon and Kelly-Riley 2004), permits an assessment of overall

critical thinking skills. Of particular interest to the subject at hand, however, is the breakdown of critical thinking scores into subcategories – including one that explicitly grades the student's ability to identify what the issue being asked about is. Over the four year span, 49 unique prompts phrasings were assigned for 31 topics[1], though in many cases the prompt varies only slightly in wording and emphasis (Appendix B).

Dependent Variables

A frustrating part of any examination process is when students simply misinterpret the question and thus fail to produce the kind of work expected – work that they might otherwise have been capable of performing. Are there ways to minimize the occurrence of such misunderstandings? Are there "types" of essay topics that students find it easier to understand? One subcategory score used as a dependent variable is the "Issue Identification and Focus Score" (hereafter "IIF score"). We would like to know whether specific types of prompts or topic wordings impact a student's ability to properly identify the issues in play and this variable is ideally suited to the task. The variable can take values between 0 and 6 in half point increments and has been normalized on a 0 to 100 scale in the study for ease of interpretation.

A second dependent variable considered in the study is the aggregate critical thinking score for an essay (Aggregate CT). This involves an average of 6 categories, also normalized to a 100 point scale. A topic might have more influence on the ability to think well than just through the ease with which students see what they have to do. Sometimes, the task is clear, but that does not make it easy. Hence, there is a need

[1] The topic in this instance refers to the distinct type of answer requested, whereas the prompt is the particular phrasing of the question. For example, "To be a good politician, must a statesman be a virtuous person? Do different principles apply to statesmen and individuals?" and "To be a good politician, must a statesman be a good person?" are the same topic but count as distinct prompt arrangements. The first prompt gives the student more information about what the topic is getting at than the second.

to see if topic and prompt effects have a wider impact on student thinking than just the IIF score. The categories scored are: issue identification and focus, logical coherence of the argument, quality and relevance of the evidence presented, assessment of the implications, ability to articulate and respond to objections, and quality of textual interpretation. (See Appendix A for more detail)

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 C.

Even if the coding rubrics are reliable enough, one might still worry about grader bias. Especially on questions of normative significance, the careful use of rubrics and grader professionalism is no guarantee that implicit bias will not creep in. As an additional check, I went through the list of topics and created an ordinal variable based on my own answers to the question "How strongly do you feel that there is a right answer to this question?" The scale went from 1 ("No Opinion on the question") to 4 (Very Strong Opinion on the question). Three results are salient: 1) There was no detectable bias for the IIF variable 2) No bias is detectable for the first three levels of intensity, but is detectable at the highest level[2]. 3) None of the results from any of the model specifications is substantially affected[3].

In addition to worries about normative bias in the assessment of critical thinking (which, to the extent that it appears, does not appear to lead to statistical bias) there are some legitimate questions about the validity of the dependent variables, especially Aggregate 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,

[2] The bias is positive, however, suggesting more lenience not less.
[3] Data and results are available upon request.

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) If we go by this definition, then the WSU measure hits close to the mark.

Methods Section

Since there is no clear literature on the subject of essay topics in political science, a combination of plausible theorizing about the likely sources of variation in critical thinking and common variables in studies of college essay scores is used to select the independent variables in the study. A step-wise regression analysis was selected as an appropriate method for identifying key explanatory variables in an exploratory study (Kelly and Maxwell 2010). First the effect size of topics as a whole (relative to other sources of variation) was tested on both independent variables. Then, a series of topic and prompt features replaced the bloc of dummy variables associated with the topics in a step-wise model to try and identify more specific sources of topic and prompt variation.

Control Variables

Control variables were chosen to account for three main sources of variation in the dependent variables: the individual characteristics of the student, the specific characteristics or context of the assignment, and varying features of the course design itself (as it changes somewhat from semester to semester).

Individual level variables: These help account for variation due to the characteristics of the students themselves. The slate of variables collected as controls was taken from the literature on critical thinking (Bataineh and Zghoul 2006; Howard, Li-Ping Tang, and Austin 2014).

Course-level variables: Although several variables could be used here to control for the effect of the semester, since semester level variables are not the variables of interest, a dummy variable for each semester was created to account for within-semester effects.

Assignment level variables: Each essay assignment also occurred in a specific class context. The first essay, typically occurring within two to four weeks of the beginning of the semester needs to be understood as taking place within a different context than the last essay, when students have between twelve and fourteen weeks of political theory under their belts. Descriptive statistics for all these variables are included in Appendix A.

Characteristics of the essay topics

Though the literature in political science is sparse, there is some discussion of essay prompts and topics in the fields of education, English as a second language (ESL), and writing. Added to variables picked from this literature were variables having to do with different question types in political theory.

Two-sidedness: Greenlaw and DeLoach (2003) argue that wording is critical to the success of online discussion questions. They suggest that the questions be "interpretive" and "allow for multiple points of view." From a slightly different angle, in the ESL context Hinkel (2002) examined prompts that provided one side of an argument versus prompts that provided two and discovered that the "two-sided" prompts engendered responses with relatively simpler contexts and lexical constructions. To operationalize the idea of multiple points of view, I created a dummy variable for whether a question was phrased in a yes/no or agree/disagree format (=1, N=543) or not (=0, N=245). Although yes or no questions can admit a multiplicity of points of view in between, perhaps the phrasing of the question in a dichotomous format can inhibit a student's ability to produce nuance by artificially reducing a continuum to a dichotomy and thus rendering the question less "interpretive."

Normativity: Political theory splits its attention between normative and empirical issues (Moore, 2011). Thus on the one hand, essay prompts sometimes ask about purely conceptual issues that do not require the student to take a normative stance and on the other, students are also asked questions that require them to defend a value judgment. Furthermore, sometimes the value judgment is explicitly requested through the use of verbs like "should" or "ought" that clearly signal that the answer needs to be normative in nature (for example: "Should governments always try to give citizens what they deserve?"). But other times, a question that requires the student to take a normative stance is phrased without the use of such verbs (for example: "What would make America great?"). It would be helpful to know whether making normativity implicit fools some students into thinking a normative question is perhaps simply conceptual or even empirical in nature. It would also be interesting to see if taking the normativity out of a question (for example by simply asking about the relationship between normative concepts without requesting the students take a stance on their importance) helps or hinders students scores. A nominal variable was therefore created coding assignments with explicitly normative topics as 1 (N=182), implicitly normative topics as 2 (N=452) and conceptual topics as 3 (N=154).

Personalization: Miller, Mitchell, and Pessoa (2007) found that among English as a Foreign Language (EFL) students, explicitly asking student for an opinion (Do you think…?) was more likely to elicit argument than asking for reasons (Why…?). Therefore, a variable coding for the degree to which the question inserts the student into the question was added. Questions that had the words "you," "your," "we," or "our" in them were coded 1 (N=337), others 0 (N=427).

Link to Reading: Hirvela and Du (2013) note that writing from source texts is difficult because it forces students to engage in "complex reading and writing activities and make contextualized decisions as they interact with the reading material and the assigned writing tasks." (87) All or nearly all the topics in the course require students to draw from one or more sets of readings. Some topics, however, were explicit about which readings they should focus on. Others required students to decide for themselves what

sources to draw upon. Any topic that explicitly mentions an author or text was coded 1 (N=192). Any topic that did not was coded 0 (N=596).

Number of Pages per Assignment: Independently of whether a reading was explicitly referenced by the prompt, the number of directly relevant readings to an essay exercise may exacerbate problems of "intertextuality." (Miller, Mitchell, and Pessoa 2007) On the one hand, the readings provide models of thinking - ways of approaching a subject. They also give students an idea of the variety of positions and types of evidence offered in favor of various views. This should be a positive factor. On the other hand, to the extent that students might be tempted to summarize someone else's view rather than offer one of their own, more readings could exacerbate the difficulty of the exercise and moreover provide more opportunities to avoid answering the question by hiding behind views they have read about rather than developing their own. This is something found among studies of writing among ESL students (Keck 2010, Pecorari 2003). A variable counting the number of pages of reading students could be expected to draw upon for each assignment was computed (Mean=106.49, SD=112.78, Range=0-443)[4].

Philosophy, Theory, or Policy: John Tomasi (2012) articulates the distinction between political philosophy, political theory, and public policy roughly as follows. All require moral judgments but each operates at a different level of abstraction. Political philosophy asks us to think about the principles and values we use to evaluate political institutions. Political theory asks us about which social and political institutions best meet our principles (once we have selected them). Public policy then asks which specific laws or regulations we should enact once we have selected the governing principles and the institutions.

[4] Admittedly, there was a judgment call in some cases about what constitutes a "relevant" reading. Clearly students fail to see the relevance of readings to certain topics over and over, since they never seem to use them but it does not follow that they couldn't have! Furthermore, although page numbers are probably more accurate than simply counting the number of readings, they do not account for differences in word count per page. As a result, the validity and reliability of this variable, depending on the editions used and density of the reading is expected to be fairly weak.

Tomasi argues that social science matters a lot at the level of policy and still quite a bit when thinking about which institutions best fit one's political values. However, facts matter decidedly less when thinking about which principles we should adopt.

We should expect political science students to be more comfortable answering questions that depend on empirical observation (theory and policy questions) than those that require purely ethical or meta-ethical reasoning. A nominal variable was created to try and capture these different types of questions students were asked to write about (Philosophy (1) = 167, Theory (2) = 318, Policy (3) = 303)

Results

Topics on the whole

Leaving aside which features of topics and prompts are important, do topic and prompts matter in the aggregate? On this question the answer seems to be yes. Using a dummy variable for each distinct topic, and considering the effect size for all the topics together (dividing the sum of squares for the topics by the total sum of squares) we see that 8.6% of the measured variation in IIF scores is predicted by variation in topics (Table 2). This is the largest source of variation, followed by individual level variation (Age, GPA, ACT scores, and Reading Percentage together accounting for 5.8%) and semester-level variation (5.3%).

Table 2
Stepwise Backward Elimination Regression Results: IIF & Aggregate CT

	IIF			Aggregate CT		
	B	SE	(η^2/r^2)	B	SE	(η^2/r^2)
Intercept	-28.90**	13.79		-20.99**	8.61	
Topic ID (Prob.>F)	.036**		0.086	.228		0.051
Semester ID (Prob.>F)	.041**		0.057	.000***		0.060
Beginning GPA	7.59***	2.29	0.021	8.84***	1.43	0.060
Gender[F]	-0.79	1.23	0.001	-1.65**	0.77	0.007

Age	1.13**	0.48	0.011	0.73**	0.30	0.009
Overall Reading Percentage	0.14***	0.05	0.016	0.16***	0.03	0.040
ACT Average	1.04***	0.31	0.022	0.91***	0.19	0.035
Hours	-0.02	0.05	0.000	-0.00	0.03	0.000
Prior Essays	3.55	2.53	0.004	3.18*	1.58	0.005
Adj. R^2		.163			.322	
R^2		.257			.398	
N		429			429	
***: p<.1; **p<.05; ***:p<.01**						

When it comes to the aggregate measure of critical thinking, the bloc of topic variables is not significant but still accounts for a relatively sizeable percentage of the variation (5.1%), equal to semester effects (5.2%). GPA (10%) gains in relative importance, with ACT scores (2.4%), Age (2%), Gender (<1%) Prior Essays (<1%) also statistically significant. Though topics are not significant as a bloc, individual topics are. Further investigation reveals that topics can have contradictory effects on different subcategories of critical thinking, which might help explain why effect sizes are consequent but statistical significance for Aggregate CT is elusive. (Appendix D)

IIF Scores

The results for IIF scores are reported in Table 4 (below).

Table 4
Stepwise Backward Elimination Regression Results: IIF scores

	All Regressors			Selected Model		
	B	SE	(η^2/r^2)	B	SE	(η^2/r^2)
Intercept	-13.48	13.49		-5.43	8.91	
Semester ID (Prob.>F)	.122		0.04	.014**		0.05
Beginning GPA	5.65***	2.31	0.01	5.45***	2.22	0.01
Gender[F]	0.62	1.25	0.00			

Age	0.67	0.45	0.00			
Overall Reading Percentage	0.15***	0.05	0.02	0.17***	0.05	0.02
ACT Average	1.10***	0.31	0.03	1.15***	0.31	0.03
Conceptual Analysis	4.50	3.32	0.01	5.84**	2.38	0.02
Hidden Normativity	-.037	2.63		-3.04*	1.79	
Philosophy	-5.77**	2.94	0.00			
Theory	-2.39	3.07	0.00			
Yes or No[Yes]	-1.29	1.88	0.01			
Personalized[Yes]	1.72	1.37	0.01			
Linked to Reading[Yes]	-4.78**	2.15	0.00	-4.71***	1.79	0.01
Prompt Length	0.29**	0.14	0.04	0.23*	0.13	0.01
Relevant pages	-0.02	0.02	0.01			
Adj. R^2		.146			.149	
R^2		.203			.190	
N		403			420	
***: p<.1; **p<.05; ***:p<.01**						

Aggregate CT

The results for the measure of Aggregate CT are presented below in Table 5:

Table 5
Regression Results: Aggregate CT

	All Regressors			Selected Model		
	B	SE	(η^2/r^2)	B	SE	(η^2/r^2)
Intercept	-14.64*	8.69		-13.12	8.46	
Semester ID (Prob.>F)	.004***		0.05	.008***		0.05
Beginning GPA	7.98***	1.48	0.05	8.02***	1.48	0.05
Gender[F]	-1.94**	0.80	0.01	-1.98**	0.80	0.01
Age	0.55*	0.29	0.01	0.56*	0.29	0.01
Overall Reading Percentage	0.15***	0.03	0.04	0.15***	0.03	0.04
ACT Average	0.87***	0.20	0.03	0.87***	0.20	0.03

Conceptual Analysis	2.14	2.14	0.00			
Hidden Normativity	-0.07	1.69				
Philosophy	-3.92**	1.90	0.02	-3.62**	1.51	0.02
Theory	0.92	1.98		1.52	1.53	
Yes or No[Yes]	0.44	1.21	0.00			
Personalized[Yes]	1.55*	0.89	0.01	1.47*	0.87	0.01
Linked to Reading[Yes]	-2.52*	1.38	0.01			
Prompt Length	0.23**	0.09	0.01	0.16**	0.07	0.01
Relevant pages	-0.01	0.01	0.00			
Adj. R^2		.288			.289	
R^2		.366			.328	
N		403			403	
***: p<.1; **p<.05; ***:p<.01**						

Discussion

Topics and prompts matter for both a student's ability to understand what they are being asked to do and Aggregate CT scores. Longer topics, other things being equal, produce slightly better scores. This is not altogether surprising since the longer topics offer more guidance and direction. Is it then desirable for the instructor to always explain what is required of the student? This will depend on the purpose of the assignment. Sometimes, what the instructor wants to assess (or the lesson she wants to impart) has to do with conceptualizing what sort of answer a question requires. In those cases, it may be best to allow students to struggle and sometimes fail. On the other hand, if the instructor is looking for students to practice the kind of critical thinking for which understanding the topic is a must (practice at a particular kind of argument, for example, or drawing a specific inference from the material) then more explanation will help.

Controlling for prompt length, that pointing to a specific reading predicts lower IIF scores in the selected model specification (and, if one takes all the regressors together, also predicts lower Aggregate CT

scores) is more of a mystery. If we are giving the student more information, why are they doing worse? Further investigation reveals that if we look at the specific topics where an author or text is mentioned, we see different kinds. Some ask students to evaluate a text or author. Others ask students to referee a debate between two class authors. Finally, some topics (121, 111) are about assessing rival textual interpretations.

We can test to see which of these was particularly difficult by creating a series of dummy variables coding for each of these groupings. We find that it is the third type of topic (asking students to assess rival interpretations of a text) that appears to be driving the negative link between mentioning texts and IIF scores (Table 6). The first two types (assessing a text or refereeing between different authors) are statistically indistinguishable from any other topic that does not mention an author. It is impossible to know at this stage whether this is a generalizable finding. The result could be driven by the specific authors in question (Hobbes and Locke, for example), instructional style, the student population at the university, characteristics of political science majors, or some combination of the above.

Table 6
Regression Results: Types of Textual Engagement and IIF Scores

	All Regressors			Selected Model		
	B	SE	(η^2/r^2)	B	SE	(η^2/r^2)
Intercept	-21.29*	14.19		-11.48	9.88	
Semester ID (Prob.>F)	.079*		0.044			0.056
Beginning GPA	5.88***	2.31	0.014	5.49**	2.21	0.012
Gender[F]	0.63	1.24	0.001			
Age	0.71	0.45	0.005			
Overall Reading Percentage	0.14***	0.04	0.016	0.16***	0.05	0.023
ACT Average	1.08***	0.31	0.025	1.14***	0.31	0.027
Conceptual Analysis	5.61	3.43	0.006	5.65**	2.48	0.017
Hidden Normativity	-2.76	2.56		-4.22**	1.76	
Philosophy	-5.25*	2.96	0.007			

Theory	-1.58	3.00				
Yes or No[Yes]	-0.54	2.08	0.000			
Personalized[Yes]	-1.89	1.42	0.005			
Text Evaluation	-1.13	3.71		-3.45	3.27	
Referee Debate	6.98	5.40	0.016	5.79	4.85	0.020
Textual Interpretation	-13.59*	7.00		-11.35*	6.98	
Prompt Length	0.46**	0.17	0.017	0.43**	0.17	0.013
Relevant pages	-0.01	0.02	0.000			
Adj. R^2		.145			.151	
R^2		.207			.195	
N		403			420	
***: p<.1; **p<.05; ***:p<.01**						

The finding that hidden normativity in an essay prompt is more problematic for IIF scores than explicit normativity is not surprising at all. Unless the instructor wants to test whether students can recognize a normative issue and distinguish that kind of issue from conceptual or empirical questions, it seems more prudent to phrase normative questions with clear identifying verbs like "should" or "ought." This signals to students that they are entering the world of normative analysis and that the kind of argument required is one that will depend on selecting the most appropriate values – not necessarily only the best facts. For political science students (who are used to talking about cause and effect) and pre-law students (who want to find answers in case law or the constitution), recognizing when a moral argument is required (not a legal or causal one) can be challenging when their intuitive approach to politics is not ideological.

Turning to aggregate CT, political science students appear to struggle with questions having to do with moral philosophy. They appear more comfortable evaluating and comparing sets of institutions or policies. What lessons we should draw from this is unclear. Should political science majors be exposed to more ethics and meta-ethics? Or should instructors instead focus on developing existing areas of (relative) strength and avoid more fundamental questions of value? The capacity to recognize when the answer to a

question turns on a question of ethics seems very valuable, but whether political science students need to be doing pure ethics as part of their major is less clear.

Conclusion

So what can practitioners take away from this study? By themselves, the findings cannot tell us what topics we should assign. That depends on what mix of skills (or attributes or attitudes) we want to measure in our students and what kinds of papers we want to grade. When we assign more meta-ethical, briefer, more impersonal topics where questions of value are hidden behind the language of moral realism that test the ability of students to interpret specific texts or authors, we should expect a greater variety of results – with lower average CT scores and many papers of the wrong type entirely.

If we explain to students in detail what we are looking for, if we let them pick the sources they are most comfortable using, try to put them personally into a more concrete situation where a normative decision must be made, and simply ask them about the relationship between various concepts without asking them to take a normative stance, we will probably get better essays that are more on point and easier to assess. But in doing this, there is a risk that we might not be sufficiently challenging our students' ability to interpret difficult texts (since they can pick ones they find more accessible), helping them practice their ethical reasoning skills, or assessing their ability to think impartially in emotionally charged contexts.

This gives rise to several areas for future research. Are there any discernible medium to long run effects of making students struggle to interpret topics versus having them demonstrate their thought processes on topics that have been interpreted for them? Does this effect vary at all depending on the caliber of the student? Can these effects be mitigated by telling students in advance that part of the challenge is figuring out what they are being asked to do? What is certain is that when we assess written essay work, it is

always in the context of a particular topic and as such, it would behoove us to decide beforehand what kind of work we want our students to produce and adjust our topic choice and prompt formats accordingly.

Works Cited

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.

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.

Condon, William, and Diane Kelly-Riley. 2004. "WSU Guide to Rating Critical Thinking." *PsycTESTS Dataset*.

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.

Franklin, Daniel, Joseph Weinberg, and Jason Reifler. 2014. "Teaching Writing and Critical Thinking in Large Political Science Classes." *Journal of Political Science Education* 10(2): 155–65.

Golub-Smith, Marna, Clyde Reese, and Karin Steinhaus. 1993. "Topic And Topic Type Comparability On The Test Of Written English." *ETS Research Report Series* 1993(1): i-36.

Greenlaw, Steven A., and Stephen B. Deloach. 2003. "Teaching Critical Thinking with Electronic Discussion." *The Journal of Economic Education* 34(1): 36–52.

Hayward, Malcolm. 1989. "Choosing an Essay Test Question: It's More than What You Know." *Teaching English in the Two-Year College*. https://eric.ed.gov/?id=EJ400283 (June 3, 2017).

Hinkel, Eli. 2002. *Second Language Writers' Text: Linguistic and Rhetorical Features*. London: Routledge.

Hirvela, Alan, and Qian Du. 2013. “Why Am I Paraphrasing?: Undergraduate ESL Writers' Engagement with Source-Based Academic Writing and Reading.” *Journal of English for Academic Purposes* 12(2): 87–98.

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.

Huot, Brian. 1990. “Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know.” *College Composition and Communication* 41(2): 201.

Ishiyama, John T., William J. Miller, and Eszter Simon. 2016. *Handbook on Teaching and Learning in Political Science and International Relations*. Cheltenham, UK: Edward Elgar Publishing.

Kassiola, Joel. 2007. “Effective Teaching and Learning in Introductory Political Theory: It All Starts with Challenging and Engaging Assigned Readings.” *PS: Political Science & Politics* 40(04): 783–87.

Keck, Casey. 2010. “How Do University Students Attempt to Avoid Plagiarism? A Grammatical Analysis of Undergraduate Paraphrasing Strategies.” *Writing & Pedagogy* 2(2).

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.

Lee, H.-K., and C. Anderson. 2007. “Validity and Topic Generality of a Writing Performance Test.” *Language Testing* 24(3): 307–30.

Miller, Ryan T., Thomas D. Mitchell, and Silvia Pessoa. 2014. “Valued Voices: Studentsâ□™ Use of Engagement in Argumentative History Writing.” *Linguistics and Education* 28: 107–20.

Moore, Matthew J. 2011. “How (and What) Political Theorists Teach: Results of a National Survey.” *Journal of Political Science Education* 7(1): 95–128.

Oliver, Eileen I. 1995. "The Writing Quality of Seventh, Ninth, and Eleventh Graders, and College Freshmen: Does Rhetorical Specification in Writing Prompts Make a Difference?" *Research in the Teaching of English* 29(4): 422–50. http://www.jstor.org/stable/10.2307/40171258?ref=search-gateway:e1e5789d5128dba13c8d00e8fa8f20c9 (June 3, 2017).

Paquette, Laure. 2015. *Teaching Political Science to Undergraduates: Active Pedagogy for the Microchip Mind*. Warsaw: De Gruyter Open.

Pascarella, Ernest T., and Patrick T. Terenzini. 2005. *How College Affects Students: a Third Decade of Research*. San Francisco: Jossey-Bass.

Pecorari, Diane. 2003. "Good and Original: Plagiarism and Patchwriting in Academic Second-Language Writing." *Journal of Second Language Writing* 12(4): 317–45.

Pomplun, Mark, David Wright, Napoleon Oleka, and Marilyn Sudlow. 1992. "An Analysis Of English Composition Test Essay Prompts For Differential Difficulty." *ETS Research Report Series* 1992(1): i-45.

Schmidt, Diane E. 2010. *Writing in Political Science: a Practical Guide*. Boston: Longman.

Tomasi, John. 2012. *Free Market Fairness*. Princeton, NJ: Princeton University Press.

Appendix A: Descriptive Statistics

Table A1
Descriptive Statistics for the Control Variables

	Mean	Std. Dev	N
ACT Average	22.7	3.95	439
GPA	3.08	0.64	781
Age	25	7.12	793
Hours	87.3	32.4	788
Overall Reading Percentage	80.2	26.2	782
Prior Essays	1.12	1.02	798

Appendix A: WSU Critical Thinking 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:

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?	55.71	24.94	793
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?	62.47	24.29	793
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)?	52.89	20.73	793
Objections	Did the student consider prominent objections to their thesis or to their assumptions? Were such objections plausibly dealt with?	32.64	26.52	793
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?	64.81	25.30	793
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?	71.74	22.76	793
Aggregate CT	An additive measure combining 6 dimensions of critical thinking (each being weighted equally).	56.64	17.15	793

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 clearly 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 and domain much cherished by political theorists. Combining Issue Identification and Identifying Context and Assumption was done because in normative contexts it is hard to separate figuring out 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 a single distinct source of variation accounted 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)

Appendix B: Topics and Prompts

Table 1
Topics by Semester and Frequency Distribution

Topic #	N	Semester	Wording
101	32	20164	Can capitalism be just?
101		20172	
102	17	20164	What would it mean for America to be "great"?
103	77	20131	Socrates' commitment to justice led to his death by public execution. Was Socrates a fool to think he must accept that fate?
103		20132	
103		20164	
103		20141	Socrates' commitment to political authority and the ideal of a just man led to his death by public execution. Was Socrates a fool to think he must accept that fate?
103		20142	
103		20154	
104	42	20142	Should governments in wealthier countries work to prevent suffering in the least advantaged countries?
104		20154	Do individuals living under the world's poverty threshold ($1.25/day) have a right to assistance from those in the world who have more than enough to meet their needs? (This, of course, includes relatively impoverished Americans) Why or why not?
104		20162	Those living in poverty at home and abroad frequently need our help. But do we have a duty to help them? And if so, why?
105	48	20124	Mill says the only thing that can warrant infringing a person's liberty is harm to another: Do you agree? Why or why not.
105		20131	
105		20132	
105		20154	Should governments prohibit adults from engaging in activities that are highly likely to cause harm to themselves? (For example: taking out a loan they are unlikely to be able to repay, use dangerous recreational drugs, drive without a seatbelt, ride a motorcycle without a helmet, self-medicate, smoke tobacco, offer themselves for prostitution, duel, etc.)
105		20162	
106	37	20154	If you knew you were going to grow up poor, would it be better to do so in a capitalist or a socialist economic system? Explain.
106		20162	
107	77	20141	Should the government be in the business of trying to make you happy?
107		20144	Is it the primary aim of good government to keep the citizens happy?
107		20152	Is the highest purpose of government to cater to the happiness of its citizens?
107		20154	Are countries good places to live in proportion to how happy their citizens are?
107		20174	Should we evaluate political systems by how happy the citizens living in them are?
108	9	20152	What is a citizen to do when manmade law conflicts with his or her understanding of what justice requires?
109	15	20144	Should government criminalize the use of recreational drugs?
110	62	20131	Does the US Federal government have authority over you?
110		20132	
110		20134	Does the United States federal government have authority over most American citizens?
110		20141	Does the US Federal Government have authority over most of its citizens? (If you decide you want to write about this topic, it's a good idea to read ahead for Monday and Tuesday)
110		20142	Does the US Federal Government have authority over most of its citizens?
110		20144	Does the United States Federal Government have authority over you?
111	16	20152	What does Locke mean when he says that appropriation from the commons is justified "at least where there is enough and as good left in common for others?" What counts

			as leaving "enough and as good" in the commons? What duties does Locke think this imposes on property owners? What duties might it impose on the state?
112	70	20133	Should the state try to make you a better person?
112		20134	Is it the government's job to try and help you become a better person?
112		20141	Should the government be in the business of trying to make you a better person?
112		20142	Is it the government's job to try and help you become a better person?
113	33	20134	Does democracy protect freedom or erode it?
113		20142	When we disagree about what rights we have, should we give the benefit of the doubt to freedom or democracy?
114	22	20132	Are there natural rights the government should respect?
114		20134	
115	32	20124	To be a good politician, must a statesman be a virtuous person? Do different principles apply to statesmen and individuals?
115		20131	To be a good politician, must a statesman be a good person?
115		20132	
115		20142	Aristotle believes that to be a good statesman you have to be a good person. Machiavelli appears to take the opposite view: a completely good person is likely to make a bad statesman. What do you think? What is the relationship (if any) between good character and good statesmanship?
116	3	20142	"Equal freedom" is an attractive ideal – endorsed by both conservatives and liberals. Yet at the same time, few can agree on what equal freedom means. Should we endorse some version of a commitment to "equal freedom"? And if so, which one?
117	24	20142	John Rawls attaches little weight to desert in his theory of distributive justice. David Schmidtz thinks desert is central to questions of distributive justice. What do you think? Should we give up on desert as Rawls suggests, or should governments consider what people deserve when thinking about what people are entitled to?
117		20174	Is there room for desert in a theory of distributive justice?
118	27	20133	Does a concern for political equality mean that we should treat individuals as political equals or does it mean we should try to make them more equal (by equalizing their life chances, certain opportunities, or certain outcomes)? Remember, you cannot (always) have both. So you must either choose or find a way to reconcile the opposing values.
118		20154	Should our governments take steps to make economic opportunities more equal?
119	2	20124	[Student selected topic]
120	10	20124	[Student selected topic]
121	49	20124	Is the Hobbesian sovereign the agent or the master of his people? Are the people now on Capitol Hill your agents, or are they your masters?
121		20131	
121		20132	
121		20152	Are the politicians on Capitol Hill our masters or our servants?
121		20172	Are your representatives on Capitol Hill your masters or your servants?
122	5	20131	Is wage labor harmful to the working class as Marx claims?
122		20132	
123	24	20152	Write an essay on global justice either addressing the claims of Kant, Beitz, or Carens. Make sure you have a single clear thesis about one or more of the claims they advance.
124	0	20152	Should governments work to reduce inequalities of income?
125	14	20152	Can you have a theory of justice without a theory of politics?
126	16	20172	Which conception of equality is more attractive: Schmidtz's or Cohen's and why?
127	9	20154	How do Hobbes and Locke's respective accounts of the state of nature influence their thinking on the nature and purposes of government?
128	3	20154	Is Socialism necessarily authoritarian?
129	12	20144	On balance, does private property enhance or diminish individual liberty?
131	13	20174	Okin says "The gendered family radically limits the equality of opportunity of women and girls of all classes." Do you agree?

Appendix C: Intercoder Reliability

Methodology

A post-test study of intercoder reliability was conducted using two senior undergraduates applying to graduate school in political science and the original grader. 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
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**

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: Individual Topic Coefficients

Table C1
Topic Coefficients and Aggregate Effect Sizes

	Focus		**Interpretation**		**Thesis**		**Objections**	
	B	SE	B	SE	B	SE	B	SE
TopicID[101]	-3.41	9.57	2.73	9.34	-13.67	9.92	2.20	10.10
TopicID[102]	-1.94	10.68	0.45	10.43	-4.40	11.08	-22.01**	11.27
TopicID[103]	-7.63	5.73	9.13*	5.60	-17.03***	5.94	-3.45	6.05
TopicID[104]	7.57	6.43	-7.51	6.27	-0.38	6.66	0.66	6.78
TopicID[105]	5.66	5.96	-3.92	5.82	2.67	6.18	4.31	6.30
TopicID[106]	3.84	7.32	-14.39**	7.15	11.37	7.59	9.18	7.73

TopicID[107]	-2.82	5.44	18.36***	5.31	-0.66	5.64	-3.24	5.74
TopicID[108]	16.65	10.56	4.54	10.32	1.33	10.96	6.81	11.15
TopicID[109]	-6.55	9.15	-3.90	8.93	6.15	9.49	23.47**	9.66
TopicID[110]	-13.49**	5.91	-8.83	5.77	-0.41	6.13	5.61	6.24
TopicID[111]	-3.99	11.31	7.40	11.05	23.01**	11.73	-24.42**	11.94
TopicID[112]	1.77	6.82	-2.19	6.66	-11.35	7.07	4.44	7.20
TopicID[113]	-5.71	7.32	-13.28	7.15	-3.53	7.59	-5.45	7.73
TopicID[114]	-7.43	8.37	-10.12	8.17	8.44	8.68	7.46	8.83
TopicID[115]	10.22	6.43	3.79	6.28	2.53	6.67	-0.70	6.79
TopicID[117]	-10.19	16.20	-4.30	15.82	7.89	16.80	31.24*	17.10
TopicID[120]	-5.32	11.31	-10.00	11.05	-8.23	11.73	14.68	11.94
TopicID[122]	6.02	6.21	-11.75	6.07	1.20	6.44	1.33	6.56
TopicID[121]	-8.33	13.10	-19.63**	12.79	4.92	13.58	-31.62**	13.83
TopicID[123]	-23.47**	10.30	-4.02	10.06	2.43	10.69	-18.08*	10.88
TopicID[125]	-15.31	10.61	-8.23	10.36	-0.95	11.00	-7.58	11.20
TopicID[126]	1.16	12.37	10.03	12.08	-20.74	12.83	-34.58***	13.06
TopicID[127]	13.45	9.89	24.12***	9.66	-3.62	10.26	-23.09**	10.44
TopicID[128]	15.51	16.21	-18.30	15.83	9.67	16.82	27.60	17.12
TopicID[129]	-13.94	11.33	-7.39	11.06	-9.97	11.75	3.16	11.96
TopicID[130]	5.04	11.90	22.58	11.62	-12.27	12.34	5.48	12.56
η^2	.085		.086		.072		.117	

Table C1
Topic Coefficients and Effect Sizes (continued)

	Evidence		Conclusions		Aggregate CT	
	B	SE	B	SE	B	SE
TopicID[101]	-20.20***	7.83	-9.03	8.78	-7.93	5.97
TopicID[102]	-12.38	8.74	-7.02	9.80	-8.58	6.67
TopicID[103]	-10.03**	4.69	-12.19**	5.26	-7.20**	3.58
TopicID[104]	-0.98	5.26	6.96	5.90	0.94	4.01
TopicID[105]	7.68	4.88	2.56	5.47	3.16	3.72
TopicID[106]	0.34	5.99	12.81*	6.72	3.94	4.57
TopicID[107]	7.32	4.45	2.26	4.99	3.46	3.40
TopicID[108]	10.41	8.65	-2.04	9.70	6.47	6.60
TopicID[109]	-10.39	7.49	-13.29	8.40	-0.74	5.72
TopicID[110]	2.27	4.84	-1.22	5.43	-2.60	3.69
TopicID[111]	21.21**	9.26	-1.76	10.38	3.88	7.06
TopicID[112]	-9.21*	5.58	5.86	6.26	-1.96	4.26
TopicID[113]	-1.87	5.99	7.01	6.72	-3.80	4.57
TopicID[114]	0.97	6.85	7.91	7.68	1.13	5.22
TopicID[115]	5.82	5.26	-4.31	5.90	2.87	4.01
TopicID[117]	-3.50	9.26	-6.87	10.38	8.60	10.11
TopicID[120]	1.78	5.09	-3.23	5.70	-3.03	7.01

TopicID[122]	-21.37**	10.72	4.54	12.02	-0.39	3.88
TopicID[121]	14.73*	8.43	-8.09	9.46	-12.09	8.18
TopicID[123]	3.52	8.68	-9.61	9.74	-5.63	6.43
TopicID[125]	-18.07	10.13	-5.56	11.35	-5.97	6.62
TopicID[126]	5.08	8.09	-10.49	9.08	-10.65	7.72
TopicID[127]	9.09	13.27	3.73	14.88	0.86	6.17
TopicID[128]	5.18	9.28	5.87	10.40	7.77	10.12
TopicID[129]	0.88	9.74	-11.22	10.92	-2.61	7.07
TopicID[130]	-3.50	9.26	-6.87	10.38	1.59	7.43
η^2	.073		.066		.051	