LAW STUDENT LEARNING GAINS PRODUCED BY A WRITING ASSIGNMENT AND INSTRUCTOR FEEDBACK

David J. Herring and Collin Lynch*

I. INTRODUCTION

By many accounts, legal education changes slowly.[1] But the pace of change has accelerated in recent years. Law schools have engaged in several forms of curricular reform.[2] In addition, some legal educators have implemented significant changes in teaching approaches.[3] Departures from the traditional large class case-

 David J. Herring is Dean and Professor of Law at the University of New Mexico School of Law. Collin Lynch is a post-doctoral researcher in the Center for Educational Informatics within the Department of Computer Science at North Carolina State University. The Authors are very grateful for the contributions, support, and guidance generously provided by Kevin Ashley, Dorothy Evensen, Thomas Ross, Ann Sinsheimer, and Lu-in Wang.

1. *See e.g.* William M. Sullivan et al., *Educating Lawyers: Preparation for the Profession of Law* 189–191 (Jossey-Bass 2007) [hereinafter *Carnegie Report*]; Philip C. Kissam, *The Ideology of the Case Method/Final Examination Law School*, 70 U. Cin. L. Rev. 137, 137–139 (2001); Edward Rubin, *What's Wrong with Langdell's Method, and What to Do About It*, 60 Vand. L. Rev. 609, 610–615 (2007); Louis N. Schulze Jr., *Alternative Justifications for Academic Support II: How "Academic Support across the Curriculum" Helps Meet the Goals of the Carnegie Report and Best Practices*, 40 Cap. U. L. Rev. 1, 1–4 (2012).

2. *See generally Carnegie Report*, *supra* n. 1, at 35–43 (describing the lawyering seminar at The City University of New York Law School and the lawyering program at New York University School of Law); Leg. Educ. Analysis & Reform Network (LEARN), *General Description of Planned Projects, 2009–2010*, at 7–14 (2009–2010) (available at http://www.albanylaw.edu/media/user/celt/learnprojects.pdf) (LEARN is a group of ten law schools that promote innovation in curriculum, teaching methods, and learning assessment.); Jessica Berg & Joanthan Adler, *The Future of Legal Education at Case Western Reserve University: Foundation, Focus, Fusion,* 95 In Brief (Summer/Fall 2013) (available at http://law.case.edu/Alumni/InBrief/Articles/TabId/799/ArtMID/1766/ArticleID/413/Reinventing-Legal-Education.aspx) (describing Case Western Reserve University School of Law's new curriculum that implements the teaching of legal theory and policy, legal doctrine, lawyering skills, and professional identity through a combination of traditional classroom methods and experiential learning); Harv. L. Sch., *HLS Faculty Unanimously Approves First-Year Curricular Reform*, http://today.law.harvard.edu/hls-faculty-unanimously-approves-first-year-curricular-reform/ (Oct. 6, 2006) (describing the reform of the Harvard Law School first-year curriculum, which, in part, requires students to complete a course in legislation and regulation and a course that addresses the impact of globalization on legal issues and systems).

3. *See e.g.* Christine Pedigo Bartholomew & Johanna Oreskovic, *Normalizing Trepi-*

dialogue teaching method followed by a single essay exam are numerous and varied. New educational interventions such as mid-term writing assignments accompanied by formative feedback, problem-based learning exercises, interactive computer-based assignments, role plays, reflective journals, and various assignments that require students to collaborate in small groups have proliferated.[4] The result is a great deal of experimentation within law schools today.

This creative movement in law teaching poses a challenge to the proponents of change. They are called to establish that the new approaches are more, or at least as, effective as the traditional approach alone. This is a difficult endeavor in such an ill-defined learning domain as the law.[5]

Meeting this challenge should proceed by first measuring the student learning gains achieved by traditional case-dialogue teaching methodology to establish a baseline for the comparison of new educational interventions. Fortunately, several legal educators and education researchers have begun to collaborate in this endeavor.[6] Their work has focused on a core goal of legal education: establishing and improving students' skill of cross-case legal reasoning.[7] To date, the researchers have found that, while traditional teaching methods do not diminish students' skill of legal

---

dation & Anxiety, 48 Duq. L. Rev. 349, 372–382 (2010) (describing best practices for legal research and writing courses).; Schulze, *supra* n. 1, at 31–41 (describing innovative academic support programs and approaches at various law schools).

4. *See e.g. Carnegie Report*, *supra* n. 1, at 35–43; Andrea A. Curcio et al., *Does Practice Make Perfect? An Empirical Examination of the Impact of Practice Essays on Essay Exam Performance*, 35 Fla. St. U. L. Rev. 271 (2008); Carol Springer Sargent & Andrea A. Curcio, *Empirical Evidence that Formative Assessments Improve Final Exams*, 61 J. Leg. Educ. 379 (2012); Schulze, *supra* n. 1, at 9–19.

5. *See generally* Kevin D. Ashley, *Teaching a Process Model of Legal Argument with Hypotheticals*, 17 Artificial Intelligence & L. 321 (2009); Judith Welch Wegner, *Reframing Legal Education's "Wicked Problems"*, 61 Rutgers L. Rev. 867, 919–923 (2009).

6. *See e.g.* Ashley, *supra* n. 5; David J. Herring & Collin Lynch, *Teaching Skills of Legal Analysis: Does the Emperor Have Any Clothes?* 18 Leg. Writing 85 (2012) (discussing the Authors' study, which will be referred to as the "Emperor's Clothes study"); Sargent & Curcio, *supra* n. 4; Dorothy H. Evensen et al., *Developing an Assessment of First-Year Law Students' Critical Case Reading & Reasoning Ability: Phase 2* (L. Sch. Admis. Council Grants Rpt. 08–02, Mar. 2008) (available at http://www.lsac.org/docs/default-source/research-(lsac-resources)/gr-08-02.pdf).

7. This form of legal reasoning calls on students to read legal texts rigorously, compare closely related cases and hypotheticals, detect indeterminacies of meaning across cases and hypotheticals, and in the end, distinguish more from less relevant comparisons. *See* Evensen et al., *supra* n. 6, at 4.

reasoning, these methods fail to produce any significant learning gains.[8]

Our initial study of learning gains in the area of cross-case reasoning provides an example of this work.[9] We focused on one six-week unit in a first-semester, first-year course.[10] The instructor exclusively used the traditional case-dialogue method of teaching, with an express goal of improving students' cross-case reasoning skill.[11] Seventy-one students began the study by taking a pre-test. The students then received six weeks of in-class instruction in the subject area of personal jurisdiction, followed by a post-test.[12] We found that,

> [o]verall, there was no significant positive movement in the development of reasoning skills once the students' post-test performance was examined relative to how much they could potentially improve based on their benchmark pre-test scores. Thus, while some students appeared to gain from their classroom experiences, these gains were not evenly or widely shared by the group as a whole.[13]

One educational intervention that has received attention from legal educators is the assignment of a mid-term writing exercise accompanied by formative feedback from the instructor.[14] The current study introduces such a writing assignment—one that was designed to engage students in cross-case reasoning. The study's hypothesis is that this intervention will produce positive learning gains in legal reasoning skill for the group as a whole.

In the end, this study extends the prior studies of traditional law teaching that failed to find any significant learning gains for the student group as a whole. Our findings indicate that by supplementing the traditional teaching method through the introduction of a short writing assignment accompanied by formative

---

8. *See id.* at 15–16, 27; Herring & Lynch, *supra* n. 6.
9. Herring & Lynch, *supra* n. 6.
10. *Id.* at 100–103.
11. *Id.*
12. *Id.*
13. *Id.* at 114.
14. *See e.g. Carnegie Report*, *supra* n. 1, at 171; Andrea A. Curcio et al., *Developing an Empirical Model to Test Whether Required Writing Exercises or Other Changes in Large-Section Law Class Teaching Methodologies Result in Improved Exam Performance*, 57 J. Leg. Educ. 195 (2007); Sargent & Curcio, *supra* n. 4, at 379.

feedback, a legal educator can produce significant, widely shared learning gains in cross-case legal reasoning skills.

II. METHODS

This study was conducted at the University of Pittsburgh School of Law, which, at the time, enrolled a first-year class of approximately 240 students. In terms of LSAT[15] scores and undergraduate GPAs,[16] student qualifications are typical for law schools that rank in the second tier.[17]

The school provides a traditional first-year curriculum. First-year students are randomly assigned to one of three sections, and every section is required to take the same set of substantive law courses.[18] Students in each section have the same instructors for their doctrinal/substantive law courses.

This study involved a single section of the first-semester Legal Process course (Civil Procedure I) taught by the lead author, David Herring, an experienced legal educator. The section contained a total of 87 students, 86 of whom completed the course. Because students are randomly assigned to sections, the class population is an adequate random sample of the incoming law students at the law school. The study assigned each participating student to one of two groups: Group 1 and Group 2.

As stated above, our goal was to assess the impact of a writing assignment accompanied by formative feedback on students' ability to engage in cross-case reasoning. In our Emperor's Clothes study,[19] we developed and tested a pair of multiple-choice tests designed to assess students' cross-case reasoning skill in the context of two personal jurisdiction cases argued before the United States Supreme Court: *Keeton v. Hustler Magazine, Inc.*[20] and

---

15. For the experimental section of 87 students, the LSAT score range was 153–168, with a mean score of 160.

16. For the experimental section of 87 students, the undergraduate GPA range was 2.4–4.0, with a mean of 3.6.

17. *See* U.S. News & World Rep., *Best Law Schools for 2013*, http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-law-schools/law-rankings (accessed Feb. 9, 2014) For the relevant year, the second tier consisted of schools that ranked 53 to 98, with four schools tied at 98. *Id.* The University of Pittsburgh ranked 91 that year. *Id.*

18. The substantive law courses for the fall semester consist of Contracts, Criminal Law, Legal Process (Civil Procedure I), and Torts.

19. Herring & Lynch, *supra* n. 6, at 100–105.

20. 465 U.S. 770 (1984).

Calder v. Jones.[21] The tests asked students to identify the legal issues at work in the cases and respond to hypothetical cases through cross-case analogies and distinctions. In that study we established that, despite the use of two different cases, the two tests were of comparable difficulty and were viable measurements of student aptitude.[22] For the purposes of this study, we developed a third multiple-choice test with the same question structure as the original tests but based on *Burnham v. Superior Court of California*.[23]

Our goals for the writing assignment were to have a clear focus on the same skills as the multiple-choice tests (i.e., close reading and cross-case reasoning) and to provide for formative feedback on students' written work product. We therefore developed a writing assignment in which students were asked to reread the *Keeton v. Hustler* oral argument and write essays that assessed each of the multiple choice responses for questions 4, 7, and 8 from the *Keeton v. Hustler* test.[24] Question 4 asked students to analyze a hypothetical case posed by one of the Justices and to

---

21. 465 U.S. 783 (1984).
22. Herring & Lynch, *supra* n. 6, at 114.
23. 495 U.S. 604 (1990).
24. The instructions for the mid-term writing assignment provided the following:

> For this exercise, you will work in a team of two. You and your partner must analyze three questions from the pre-test based on the oral argument in *Keeton v. Hustler Magazine, Inc.* that you completed earlier this semester. The pre-test is posted on the course's TWEN website. This is the only document that you should use in completing this exercise.
>
> You must address questions 4, 7, and 8 from the pre-test.
>
> For question 4, you must explain why responses a. and d. are correct and why b. and c. are incorrect.
> For question 7, you must explain why responses c. and d. are correct and why a., b., and e. are incorrect.
> For question 8, you must explain why responses a., c., and f. are correct and why b., d., e., and g. are incorrect.
> Your explanations must be logical and concise. You must limit your paper to a total of five double-spaced typed pages. You must not discuss this exercise with anyone other than your partner. You must submit your paper to the Registrar's Office by 3 p.m. on Tuesday, October 12, 2010.
>
> You and your partner will receive one grade (outstanding, excellent, adequate, inadequate) and written feedback that is intended to assist you in acquiring basic skills of legal analysis and in preparing for the final exam in this course. Your performance on this writing exercise will constitute 10% of your grade for the course.

select all of the responses that appropriately characterized the hypothetical's relationship to the legal test being proposed by the plaintiff's attorney.[25] Question 7 asked students to select all of the responses that appropriately characterized the attorney's response to a second hypothetical case, and question 8 asked the students to select all of the hypothetical cases that would pose a challenge to the attorney's proposed legal test.[26]

Students worked on the writing assignment in teams of two as assigned by the instructor, with both Group 1 and Group 2 having one team of three. The work was expected to be collaborative for all of the questions, and the students were graded as a team. For all three questions, the writing assignment required

---

25. For illustrative purposes, question 4 on the *Keeton v. Hustler* test is set out below:

> Assume that Mr. Grutman's proposed test is as follows: "If the state long-arm statute is satisfied and defendant has engaged in purposeful conduct directed at the forum state out of which conduct the cause of action arises, and that conduct satisfies the minimum contacts under which substantial justice and fair play make it reasonable to hail defendant into court there, and the forum state has an interest in providing a forum to the plaintiff, then the forum has personal jurisdiction over the defendant for that cause of action."
>
> The following hypothetical was or could have been posed in the oral argument. It is followed by some explanations why the hypothetical is or is not problematic for Mr. Grutman's proposed test.
>
> Please check ALL of the explanations that are plausible.
>
> Hypothetical: "Just to clarify the point, that would be even if the plaintiff was totally unknown in the jurisdiction before the magazine was circulated?" [i.e., suppose the plaintiff was totally unknown in the state before the magazine was circulated. Would personal jurisdiction over Hustler Magazine lie in that state?]
>
> a. The hypothetical is problematic for Mr. Grutman's proposed test. The decision rule applies by its terms, but arguably the publisher should not be subject to personal jurisdiction in the state under those circumstances.
> b. The hypothetical is not problematic for Mr. Grutman's proposed test. The decision rule applies by its terms, and the publisher should be subject to personal jurisdiction in the state under those circumstances.
> c. The hypothetical is problematic for Mr. Grutman's proposed test. The decision rule does not apply by its terms, but arguably the publisher should be subject to personal jurisdiction in the state under those circumstances.
> d. The hypothetical is problematic for Mr. Grutman's proposed test. The decision rule applies by its terms, but publishers would then be subject to personal jurisdiction even in a state where plaintiff suffered no injury. Include the answers to this and the following questions?

26. All of the test instruments used in this study are on file with the first Author (Herring) and are available upon request.

each student team to explain why each of the multiple choice responses was considered either correct or incorrect. Specifically, the assignment did not ask students to compose and explain their own responses, but instead required them to provide the reasoning for the responses provided by the instructor. Following completion of the writing assignment, students received formative feedback in the form of a standardized rubric that provided a sample answer,[27] along with individualized comments from the instructor on the students' answers that included specific critiques of their reasoning and suggestions for improvement.

In this study, we made use of a within-subjects design wherein all of the students experienced the same educational interventions by the end of the study period, although, as described below, not at the same time, and each student completed all three tests—pre-test, mid-test, and post-test.[28] The use of a within-

27. For illustrative purposes, the formative feedback rubric related to question 4 on the *Keeton v. Hustler* test is set out below:

> **Question 4** (Explain why a. and d. are correct; b. and c. are incorrect—responses intended by test designers)
>
> a. Correct because it is plausible to view defendant having engaged in purposeful conduct directed at the forum state (circulated magazine in the state), to view this conduct as giving rise to the cause of action (defamatory statement appears in the magazine), to view this purposeful conduct as satisfying the minimum contacts test (purposeful availment), and to view the forum state as having an interest in providing a forum to the plaintiff (New Hampshire has an interest in prohibiting false statements from being distributed to state residents). Because the decision rule plausibly applies by its terms, the proposed test indicates that there is personal jurisdiction over the defendant. This is problematic because the plaintiff has not sustained an injury to her reputation in the forum, and therefore, has a non-existent or very low interest in litigating in the forum that arguably should preclude the forum state from exercising personal jurisdiction over the defendant.
> d. Correct because of the same basic reasoning as a. above. The difference is that the response here provides an express reason for concern with personal jurisdiction in this case.
> b. Incorrect because it is not plausible to argue that plaintiff has a sufficient interest in litigating in the forum that allows the forum state to exercise personal jurisdiction over the defendant.
> c. Incorrect because it is not plausible that the decision rule does not apply by its terms. The forum state's interest is rather weak, but it clearly exists as indicated in a. above.

28. The within-subjects design allowed us to compare the test performances of each individual student over the course of the study period. It should be noted that, in part, the purpose of this design was to satisfy Institutional Review Board (IRB) requirements that the in-class experiment not affect or bias the students' final course grades.

subjects design eliminated variations in class instruction as a factor in student performance because both groups experienced the same in-class lectures and study environment. Students were randomly assigned to two groups using balanced random assignment based upon their small section Legal Analysis and Writing assignment[29] and their performance on the pre-test. In this structure students are ranked by their pre-test score and then randomly assigned to each group by score.

Initially, all students completed the *Keeton v. Hustler* test in class. Both groups then attended the same class discussions for the next four weeks. Students in Group 1 were then tasked with completing the *Keeton v. Hustler* writing assignment while students in Group 2 were given a small group problem-solving exercise that culminated in a traditional class discussion, with the exercise designed so that students would spend approximately the same amount of time on the exercise as students would spend on the writing assignment. Group 1 completed the writing assignment over a three-day period. The class sessions then continued for two weeks, with only Group 1 receiving written formative feedback during this two-week period. Both groups then took the *Calder v. Jones* test (mid-test) in class. Following the mid-test, the class sessions continued, with the class discussing issues other than personal jurisdiction. Students in Group 2 completed the *Keeton v. Hustler* writing assignment two weeks after the mid-test while Group 1 completed the small group problem-solving/class discussion exercise. Students in Group 2 received formative feedback the following week and both groups then took the *Burnham* test in class (post-test).

A timeline of these events is set out below:

Week 1:

- **Pre-test** (all students)
- Class discussions of personal jurisdiction cases

29. In addition to their substantive law courses, students must also take a Legal Analysis and Writing course, for which each section is divided into three small sections, with each small section having a separate instructor.

Weeks 2–4:

- Class discussions of personal jurisdiction cases

Week 5:

- Mid-term writing assignment (Group 1 students)
- Small group problem-solving exercise (Group 2 students)
- Class discussions of personal jurisdiction cases

Week 6:

- Formative feedback on mid-term writing assignments (Group 1 students)
- Class discussions of personal jurisdiction cases

Week 7:

- **Mid-test** (all students)
- Class discussions of notice and opportunity to be heard cases (all discussions of personal jurisdiction cases completed)

Week 8:

- Class discussions of federal subject matter jurisdiction cases

Week 9:

- Mid-term writing assignment (Group 2 students)
- Small group problem-solving exercise (Group 1 students)
- Class discussions of federal subject matter jurisdiction cases

Week 10:

- Formative feedback on mid-term writing assignment (Group 2 students)
- Class discussions of federal venue and forum non conveniens cases

Week 11:

- **Post-test** (all students)
- Class discussions of *Erie* doctrine cases
- (Study period complete)

The intent of this design was to have one group of students (Group 1) complete the writing assignment while they were in the midst of the personal jurisdiction unit of the course and to have another group of students (Group 2) complete the writing assignment following the completion of the personal jurisdiction unit. This design allowed us to measure the effects, if any, of the timing of the writing assignment.

For the purposes of this study, our analysis of the students' performance is based primarily on the three test scores. We assessed the impact of the writing assignment using the pre- and mid-test scores for both groups, and we assessed the long-term impact of the writing assignment and feedback, as well as the impact of the timing of the writing assignment by examining the post-test scores. While we examined the students' incoming competence measures such as Law School Admissions Test (LSAT) score and Undergraduate Grade Point Average (UGGPA), it is important to note that these measures assess a set of reasoning skills, only a portion of which are assessed by our tests.[30] Therefore, we expected the relative correlation of these measures with test performance to be low. Similarly, the students' final course grade was based upon a number of topics not covered during the study period. That, combined with the within-subjects design, led us to expect that the course grades would not differ between the groups or be a useful outcome measure.

III. RESULTS

A. Group Descriptions

We conducted this study using a within-subjects design. The class began with 87 students of whom 82 completed all of the study tasks. Fifty-two of the 82 students were male and 30 female. Group 1 consisted of 40 students who completed the study, while Group 2 contained 42 students. We found no significant difference between the groups in terms of their LSAT scores or UGGPAs. The gender breakdown of the groups was slightly skewed with 29 of 40 students in Group 1 being male while 23 of 42 students in Group 2 were male.

---

30. *See e.g.* Curcio et al., *supra* n. 4, at 283–286.

B. Test Results

As we stated above, the students in this study completed a series of three tests designated as pre-test, mid-test, and post-test. A comparison of the mean group test scores[31] is shown in the table below.[32]

TABLE 1: Mean Test Scores and Statistical Comparisons Using One and Two-Sided t-tests

Test	Group 1	Group 2	Analysis	P-Value
Pre-Test	3.36	3.643	Two-Sided G1 ≠ G2: t(39)=-0.81	p<0.42
Mid-Test	4.75	4.033	**Two-Sided G1 ≠ G2: t(39)=1.8**	**p<0.052**
Post-Test	5.12	4.625	Two-Sided G1 ≠ G2: t(39)=1.43	p=0.157

31. In our previous work we reported on a series of grading metrics. Herring & Lynch, supra n. 6, at 106. For the purposes of this Article, we have focused solely on the *Even* grading rubric and all the results reported here are drawn from that rubric. The *Even* grading rubric allows us to accommodate appropriately the two classes of questions used on the tests in this study (i.e., single choice questions and multi-choice questions). Single choice questions ask students to choose the one correct answer. For those questions students receive a single point for a correct choice and no points for an incorrect choice. For multi-choice questions, students receive a weighted positive score for each correct choice and lose a weighted value for each incorrect choice so that the range of possible scores is -1 to 1. On a question with two correct and three incorrect answers, for example, students receive 1/2 point for each correct choice and lose 1/3 point for each incorrect choice.

32. The comparisons listed in this table are two-sample t-tests for the inequality of the group scores. A two-sample t-test is a statistical comparison that tests for differences between two means. The p value represents the probability of a type 1 error or the probability of concluding that group 1's score differs from that of group 2 when it does not. *See* B.L. Welch, *The Generalization of "Student's" Problem When Several Different Population Variances Are Involved*, 34 Biometrika 28 (Jan. 1947).

A representation of these scores can be seen in the plot below:

Plot Representation of Test Scores

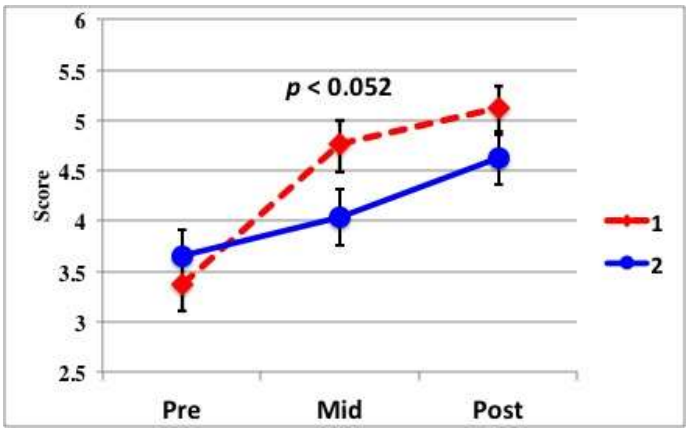

Table 1 and the Plot show that Group 1 scored higher on the mid-test than Group 2 at a marginally statistically significant level, with Group 2 making up some of the difference in their performance on the post-test.

C. Gain Scores

While the test scores are important, it is also important to examine the gain scores—that is, the relative improvement of individual students across tests. For the purposes of this study we use two gain scores—the *Raw Gain* and the *Normalized Learning Gain* (NLG). Raw Gain is simply the difference between a student's final score and his or her initial score (e.g., post-test score—pre-test score). While this gain measure is instructive, it can obscure small but important changes. Students who score very poorly on the pre-test have a large amount of potential gain. As a consequence, these students can obtain a much higher gain score than students who score relatively high on the pre-test and have less room to improve. The normalized score addresses this issue by dividing the raw gain score by the difference between the maximum possible score and the pre-test score as shown below. This normalization allows us to detect whether the population as a whole gained the same relative to their pre-test scores.

$$\frac{\text{Post-test score} - \text{Pre-test score}}{\text{Max possible score} - \text{Pre-test score}}$$

Raw Gain scores and NLG scores for both groups across all three tests are shown below in Table 2 and Table 3, respectively.

TABLE 2: Raw Gain Scores by Group with Comparisons and Power[33]

Gain Score	Group	Value	Test	P-Value	Power
Pre-Mid Raw	1	1.395	**Two-Sided m ≠ 0: t(39)=3.86**	**p<0.001**	
Mid-Post Raw	1	0.363	Two-Sided m ≠ 0: t(39)=1.21≠	p<0.231	0.22
Pre-Post Raw	1	1.750	**Two-Sided m ≠ 0: t(39)=6.71**	**p<0.001**	
Pre-Mid Raw	2	3.798	Two-Sided m ≠ 0: t(41)=1.04	p<0.303	0.02
Mid-Post Raw	2	0.592	Two-Sided m ≠ 0: t(41)=1.49	p<0.141	0.3
Pre-Post Raw	2	0.972	**Two-Sided m ≠ 0: t(41)=0.972**	**p<0.02**	

33. Power is the probability that a null result, no difference, proves that no difference exists in the population.

TABLE 3: NLG Scores by Group with Comparisons and Power

Gain Score	Group	Value	Test	P-Value	Power
Pre-Mid NLG	1	0.152	**Two-Sided m ≠ 0: t(39)=2.79**	**p<0.009**	
Mid-Post NLG	1	-0.050	Two-Sided m ≠ 0: t(39)=-0.485	p<0.631	0.06
Pre-Post NLG	1	0.235	**Two-Sided m ≠ 0: t(39)=5.45**	**p<0.001**	
Pre-Mid NLG	2	-0.010	Two-Sided m ≠ 0: t(41)=-0.14	p<0.886	0.03
Mid-Post NLG	2	-0.004	Two-Sided m ≠ 0: t(41)=-0.05	p<0.958	0.03
Pre-Post NLG	2	0.082	Two-Sided m ≠ 0: t(41)=1.25	p<0.218	0.23

Group 1 achieved significant Raw Gains from pre-test to mid-test and from pre-test to post-test. Group 1 also achieved significant NLG's from pre-test to mid-test and from pre-test to post-test. For Group 2, we found significant Raw Gains only from pre-test to post-test.

In comparing groups in terms of gain scores, we found marginally significant differences in favor of Group 1 for the pre-test to mid-test gains and for the pre-test to post-test gains. These significant differences were found for both Raw Gains and NLG's as shown below in Table 4 and Table 5.

TABLE 4: Cross-Test Naw Learning Gain Scores and Statistical Comparisons

Comparison	Group 1	Group 2	Analysis	P-Value	Power
Pre-Mid Raw	1.395	0.3798	Two-sided G1 ≠ G2: t(39)=1.98	**p<0.051**	
Mid-Post Raw	0.3628	0.592381	Two-Sided G1 ≠ G2: t(39)=-0.46	p<0.65	0.17
Pre-Post Raw	1.76	0.97	**Two-Sided G1 ≠ G2: t(39)=1.75**	**p<0.084**	

TABLE 5: Cross-Test NLG Scores and Statistical Comparisons

Comparison	Group 1	Group 2	Analysis	P-Value	Power
Pre-Mid NLG	0.15210	-0.009842	**Two-Sided G1 ≠ G2: t(39)=1.86**	**p<0.066**	
Mid-Post NLG	-0.0505	-0.00420	Two-Sided G1 ≠ G2: t(39)=-0.35	p<0.73	0.05
Pre-Post NLG	0.23	0.08	**Two-Sided G1 ≠ G2: t(39)=1.95**	**p<0.055**	

D. Demographic Comparisons

Once the study was completed, we tested for potential aptitude-treatment interactions,[34] we tested the correlation between

34. Aptitude-treatment interactions occur when students' performance on or receptiveness to an educational intervention is affected by their incoming aptitude. *See* Lee J. Cornbach & Richard E. Snow, *Aptitudes and Instructional Methods: A Handbook for Research on Interactions* (Irvington Publishers 1981). For example, better-prepared students may be more prepared for an independent writing assignment or better able to grasp the fundamental legal concepts used in the cases under discussion, whereas lower-scoring

students' incoming aptitude variables (LSAT scores and UGGPAs) and their test scores using Spearman's Rank Sum Correlation test[35] for comparisons both within and across the groups. We followed up on these comparisons by subdividing the groups based on their LSAT scores and UGGPAs (using median split), as well as their gender, to test for possible interactions.

1. Aptitude Correlations

While we found a positive correlation between the students' LSAT and UGGPA and their pre-test scores when measured across the whole group, these correlations were not significant (LSAT: $\rho(80)=0.161$, $p<0.15$, UGGPA: $\rho(80)=0.022$, $p<0.83$). Nor did we find significant correlations between their UGGPA and their mid-test or post-test scores or any of the gain scores when assessed both as a whole population and within groups.

We did, however, find a significant correlation between the students' LSAT and their post-test scores across the whole population: $\rho(80)=0.4$, $p<0.01$. We also found a positive correlation between the pre-test to post-test NLG ($\rho(80)=0.19$; $p<0.09$) and both of the mid-test to post-test gain scores: $\rho(80)=0.27$, $p<0.014$ (Raw Gain); $\rho(80)=0.29$ $p<0.01$ (NLG).

When we tested these correlations on a group-specific score we found no statistically significant or marginally significant correlations for students in Group 1. We did, however find that for students in Group 2 their LSAT was positively correlated with their post-test performance: $\rho(80)=0.54$, $p<0.01$. We further found that there were significant positive correlations with their mid-test to post-test gain scores (Raw Gain: $\rho(80)=0.33$ $p<0.03$; NLG: $\rho(80)=0.4$, $p<0.011$) and marginally significant positive correlations with their pre-test to post-test NLG ($\rho(80)=0.29$, $p<0.07$). These results suggest that the high LSAT students in Group 2 gained more from the educational experience than low LSAT students in Group 2, while this was not the case in Group 1.

---

students may respond better to more structured guidance. *See* Richard E. Snow, *Aptitude—Treatment Interaction as a Framework for Research on Individual Differences in Learning* (Ctr. for Educ. Res. at Stanford 1988).

35. See Spearman's rank sum correlation for Rho as described in C. Spearman, *The Proof and Measurement of Association between Two Things*, 15 Am. J. Psychol. 72 (1904).

2. Aptitude Subgroups

For each group, we assigned students to two subgroups by UGGPA for purposes of analysis only, using the median class score of 3.39 as the dividing line. We split Group 1 into two groups containing 16 and 24 students respectively. Group 2 was split into groups of 24 and 18 respectively. We then compared the high and low groups using a one-way analysis of variance coupled with a series of pairwise t-tests.[36] We found that the group assignment was significant with respect to their pre-test to post-test NLG scores $F(3,78)=2.8$, $p<0.05$.[37] Subsequent pairwise t-tests showed that the differences between the high-UGGPA students in Group 1 and the high-UGGPA students in Group 2 were statistically significant: Group 1 mean 0.311; Group 2 mean -0.004; pairwise p-value $p<0.043$.

For LSAT, the median class score was 160. Here we performed a similar median split segmenting Group 1 into groups of 14 and 26 respectively while Group 2 was split into groups of 14 and 28 respectively. Commensurate with the correlation that we observed above, we found a marginally significant difference between the LSAT groups on their post-test score ($F(3,78)=2.635$, $p<0.06$). Post-hoc pairwise t-tests, however, identified no statistically significant cross-group differences.

3. Gender

We first examined differences by gender on the incoming aptitude variables. We found that the women had a mean LSAT score of 156.3 with a standard deviation of 3.84. The men had a mean LSAT score of 158.36 with a standard deviation of 4.22. This difference was marginally significant according to a t-test ($t(54)=-1.86$, $p<0.068$). There was, however, no significant difference in UGGPAs. The women had a mean UGGPA of 3.41 with a standard deviation of .29; while the men had a mean UGGPA of 3.31 with a standard deviation of .37 ($t(54)=1.105$, $p<0.28$).

36. For all pairwise t-tests the Holm correction metric was employed. *See* S. Holm, *A Simple Sequentially Rejective Multiple Test Procedure*, 6 Scandinavian J. Statistics 65 (1979).

37. The F-scores presented in this study are taken from one-way or two-way analysis of variance tests (ANOVA). ANOVA tests for statistically-significant differences among a set of groups. A one-way ANOVA comparison of two groups is equivalent to a two-sided t-test.

Therefore, there were some aptitude differences by gender at the outset of the study.

We then divided students by gender both within and across the groups in order to identify any potential differences related to the study's tests. We found no significant differences between the groups in terms of their pre-test and mid-test scores. We did, however, find a difference in their post-test scores (means: male 5.281; female 4.148; $F(1,54.6)=10.17$, $p<0.003$), pre-test to post-test Raw Gains (means: male 1.744; female: 0.6823; $F(1,50.9)=4.69$, $p<0.04$) and NLGs (means: male 0.2228; female 0.0425; $F(1,50.9)=4.53$, $p<0.04$), and mid-test to post-test Raw Gain scores (male: 0.8527; female: -0.165; $F(1,55.27)=3.8$, $p<0.06$).

Within the groups we found that gender was a significant factor only for Group 2. More specifically, we found that male and female students in Group 2 had significantly different post-test scores: male mean score 5.301; female, 3.807; $F(1,38.786)=9.6$, $p<0.004$. This difference also applied to their pre-test to post-test gain scores and mid-test to post-test gain scores as shown in Table 6 below:

TABLE 6: Gender Differences in Gain Scores within Group 2

Score	Test	P-Value
mid-post Raw	$F(1,38.733)=7.51$	$p<0.01$
mid-post NLG	$F(1,28.91)=6.8$	$p<0.015$
pre-post Raw	$F(1,35.6)=3.8$	$p<0.058$
pre-post NLG	$F(1,38.79)=3.5$	$p<0.07$

While the groups are too small to confirm that the absence of a difference for Group 1 is significant, this distinction between the groups may indicate possible gender differences in the effectiveness of the particular educational interventions that we introduced in this study. These findings merit further study.

E. Final Exam and Writing Assignment

As stated above, we applied the within-subjects design with the goal of eliminating any student variation on the final exam.[38] Moreover, we ensured that students were given the same instructions and time to complete the writing assignment. Therefore, we hypothesized that the two groups would score equally on both the final exam, which measured the whole class content, and on the writing assignment.

Our findings supported these hypotheses. We found no significant difference between the groups on the writing assignment score: Group 1 mean, 8.7; Group 2, 8.6; $t(39)=0.4207$, $p<0.68$, power<0.09. Similarly, we found no significant difference between the groups on their final exam score: Group 1 mean score, 67.82; Group 2, 69.31; $t(39)=-0.87$, $p<0.38$, power>0.99.

IV. DISCUSSION

This study examined the possible benefits, in terms of student learning gains, of supplementing the traditional case-dialogue method of law teaching with one short writing assignment accompanied by formative feedback. The study also examined whether the benefits varied depending on the timing of the writing assignment, with one group of students completing the assignment during the unit of study and the other group completing the assignment following the unit of study. The study found significant benefits from the writing assignment, with students who completed the assignment during the unit benefitting more than students who completed the assignment following the unit of study.

The study involved 82 first-semester law students who were divided into two groups that were similar in terms of LSAT scores and Undergraduate Grade Point Averages (UGGPAs). Both groups began at a similar point in terms of test performance, with the pre-test scores revealing no significant difference between the groups. However, there was a significant difference between the groups on mid-test scores, with Group 1 scoring significantly higher than Group 2. In the end, there was no significant differ-

38. *See supra* nn. 28–29 and the accompanying text (providing an explanation of the study's within-subjects design).

ence between the groups in terms of post-test scores. As to learning gains, students in Group 1 achieved both significant Raw Gains and Normalized Learning Gains (NLGs) from pre-test to mid-test and from pre-test to post-test. The significantly positive NLGs indicate that the gains were widely shared by members of Group 1 as a whole. In contrast, students in Group 2 achieved only significant Raw Gains from pre-test to post-test. Thus, while some students in Group 2 appeared to gain from their educational experiences over the course of the entire study period, the gains were not evenly or widely shared among the members of Group 2 as a whole.

The finding of significant NLGs for Group 1 departs from the findings of previous studies that examined learning gains achieved by the traditional case-dialogue method of law teaching alone.[39] As discussed above, these prior studies failed to find any widely shared significant learning gains over the course of a substantive law learning unit,[40] the first year of law school,[41] or the entire three-year span of a traditional legal education.[42] It appears that the additional educational intervention of a short writing assignment accompanied by formative feedback during the unit of study produced significant, widely shared learning gains.

The differences in learning gains between Group 1 and Group 2 indicate an effect produced by the timing of the writing assignment and the accompanying formative feedback. The law students in this study who completed the writing assignment and received feedback while they were in the midst of the educational unit to which the assignment relates achieved marginally significantly greater learning gains overall than the students who completed the writing assignment and received feedback two weeks after completing the educational unit. Thus, it appears that a writing assignment will produce larger learning gains if it is given to the students while they are still involved in the relevant educational unit rather than waiting until students have completed the educational unit. This difference-in-time effect appeared to be especially strong for women students, with women in Group 2 (the group that received the writing assignment after the

39. *See* Ashley, *supra* n. 5; Evensen et al., *supra* n. 6; Herring & Lynch, *supra* n. 6..
40. Ashley, *supra* n. 5; Evensen et al., *supra* n. 6; Herring & Lynch, *supra* n. 6.
41. Evensen et al., *supra* n. 6, at 15, 27.
42. *Id.* at 16, 27.

educational unit was finished) achieving significantly lesser learning gains than men in Group 2. The effect also appeared to be strong for low LSAT students, with low LSAT students in Group 2 achieving significantly lesser learning gains than high LSAT students in Group 2. There were no similar differences noted by gender or LSAT score for students in Group 1. Thus, it appears that women and low LSAT students in Group 1 benefitted as much as men and high LSAT students from a writing assignment given during the relevant unit of study.

In terms of students' incoming qualifications, LSAT scores predicted both post-test scores and pre-test to post-test NLGs for the study group as a whole. This latter finding is interesting in light of the finding from our initial study, Emperor's Clothes, that LSAT scores failed to reliably predict learning gains produced by the traditional case-dialogue method of law teaching alone.[43] It is possible that the writing assignment intervention supports learning gains in a way that correlates with LSAT scores. In contrast to LSAT scores, UGGPAs failed to reliably predict either test performance or learning gains.

Student gender predicted post-test scores for the study group as a whole, with men scoring significantly higher than women. Gender also predicted pre-test to post-test learning gains for the study group as a whole, with men achieving significantly greater learning gains than women. It should be noted, as discussed in the "Gender" subsection above, that the men had a marginally significantly higher mean LSAT score than the women. Also as noted above, the significant gender differences arose exclusively within Group 2. These findings call for additional study concerning gender and learning gains produced by educational interventions designed to enhance legal reading and reasoning skills.

Despite the significant difference in gain scores from pre-test to post-test between the two groups in the first six-week unit of the course, there were no significant differences between the groups in performance on a final essay exam given after the fourteen-week course was completed. Although the within-subjects design of the study[44] contributed to this no-difference-in-the-end result, it is interesting that students in Group 1 overall failed to

43. *See* Herring & Lynch, *supra* n. 6, at 112–115.

44. *See supra* nn. 28–29 and the accompanying text (providing an explanation of the study's within-subjects design).

transfer their superior learning gains from the personal jurisdiction unit to their overall performance on the final exam. There are several plausible explanations for this result. It could mean that students who achieved superior gains in reasoning skills within the personal jurisdiction unit were unable to transfer their reasoning skills to other units of the course that were tested on the final exam.[45] In the alternative, it could mean that students in Group 2 were somehow able to catch up to Group 1 in terms of reasoning skills over the last eight weeks of the course. Or it could mean that the final exam failed to accurately measure the basic legal reading and reasoning skills assessed by the study's tests. The final exam called for an essay response, departing from the multiple choice format used in the study's assessment instruments. The essay examination format introduces a higher degree of subjectivity in grading and increases the opportunity for grader error and bias. In other words, it is possible that the finding of no difference between the groups on the final exam is a product of the instructor's incapacity to assess basic reading and reasoning skills in the grading of an essay. The study's data do not allow us to determine which alternative meaning explains the result here. This finding calls for further study.

This study has significant limitations. The participants were students at one law school who were all enrolled in the same set of first semester courses with the same set of instructors. In fact, the study focused on a single area of doctrine covered in one course for a period of six weeks by one instructor. The study's findings, therefore, are not immediately generalizable to other instructors or other law schools, let alone to legal education as a whole.

These limits are real and significant. Nonetheless, this pilot study presents particularistic information that has utility if taken up by researchers in different situations. In other words, this study's findings support further inquiry in this area. There is a need to continue testing new educational interventions and measuring their effects.

We plan to pursue this line of inquiry by introducing social media software that allows students to read and annotate legal

45. Other subjects covered in the course included federal subject matter jurisdiction, the applicable law in the federal courts in diversity of citizenship cases—i.e., the *Erie* doctrine.

texts together. We will measure whether this additional educational intervention, in conjunction with traditional case-dialogue and writing assignments accompanied by formative feedback, produces enhanced learning gains in terms of basic legal reading and reasoning skills. In the end, we hope to develop various evidence-based methods for achieving learning gains in this basic area of legal education.

CONCLUSION

This study indicates that the introduction of a modest educational intervention (i.e., a short writing assignment accompanied by formative feedback) within the context of a traditional law course produces significant learning gains. Thus, there appears to be hope that law teaching can produce measurable gains in novice law students' basic legal reading and reasoning skills. In this study, this hope was most fully realized for students who completed the writing assignment and received feedback while they were in the midst of struggling with the relevant substantive law material. Students who completed the writing assignment and received feedback after they had completed the educational unit achieved lesser learning gains. These findings are interesting and useful, providing support for further rigorous empirical study of law student learning.