TEACHING SKILLS OF LEGAL ANALYSIS: DOES THE EMPEROR HAVE ANY CLOTHES?

David J. Herring and Collin Lynch[*]

INTRODUCTION

Curricular reform has become a much discussed topic among legal educators. As a result, numerous law schools have considered changes such as the introduction of a legislation and regulation course in the first-year curriculum, the expansion of courses that examine international and comparative law issues, an increase in courses that address practical lawyering skills, and the development of skills or substantive pathways that provide students with a sense of progressive learning in particular areas.[1] Discussions of teaching methodologies are also prevalent. For example, many law teachers have considered how to provide stu-

[*] © 2012, David J. Herring and Collin Lynch. All rights reserved. David J. Herring is Professor of Law at the University of Pittsburgh School of Law. Collin Lynch is a Doctoral Candidate in the Learning Research and Development Center & Intelligent Systems Program, University of Pittsburgh.

We are very grateful for the contributions, support, and guidance generously provided by Kevin Ashley, Dorothy Evensen, David Harris, Thomas Ross, Ann Sinsheimer, James Stratman, and Lu-in Wang.

[1] *See generally* Case Western Reserve U. Sch. of L., *Academics*, *Experiential Learning*, http://law.case.edu/Academics/ExperientialLearning.aspx (accessed Oct. 31, 2012) (describing Case Western Reserve University School of Law's Case*Arc* Integrated Lawyering Skills Program that merges the teaching of legal theory and policy, legal doctrine, lawyering skills, and professional identity through a combination of traditional classroom methods and experiential learning); Leg. Educ. Analysis & Reform Network ("LEARN"), *General Description of Planned Projects 2009–2010*, at 7–14 (2010) (copy on file with Authors) (LEARN is a group of ten law schools that promote innovation in curriculum, teaching methods, and learning assessment.); William Mitchell College of L., *Pathways to the Profession of Law: A Comprehensive Path through Your Legal Education*, http://www.wmitchell.edu/ pathways/ (accessed Sept. 29, 2012) (describing William Mitchell College of Law's Pathways Program, an interactive web-based application that "helps students choose, sequence, and prioritize courses by showing them visual representations of different routes through the upper-level curriculum"); *Rethinking Langdell: Historic Changes in 1L Curriculum Set Stage for New Upper-Level Programs of Study*, http://www.law.harvard.edu/news/today/dec_hlt_langdell.php (accessed Sept. 29, 2012) (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).

dents with frequent and effective formative feedback and how to assess student learning outcomes.[2]

The increased attention to these aspects of legal education is in large part due to the publication of the *Carnegie Report* on legal education.[3] The *Carnegie Report* describes the development of three pillars of legal education: the development of skills of legal analysis; practical lawyering skills; and professional identity. The authors of the *Carnegie Report* are largely critical of current approaches to legal education, faulting law schools for failing to adequately prepare students for the practice of law. They find legal education to be especially lacking in developing practical skills and professional identity. In the end, the authors urge legal educators to develop an integrative model that effectively addresses each of the three pillars of legal education.[4]

Through their critique, the authors of the *Carnegie Report* assert that, at the least, law schools currently do one thing very well—use the case-dialogue method to teach legal analysis. They state,

> [L]aw schools are impressive educational institutions. In a relatively short period of time, they are able to impart a distinctive habit of thinking that forms the basis for their students' development as legal professionals. In our visits to over a dozen law schools of different types and geographical locations, our research team found unmistakable evidence of the pedagogical power of the first phase of legal education.[5]

The authors conclude that legal educators are effective in teaching students how to think like a lawyer.[6]

Based on a line of empirical studies, this Article questions the *Carnegie Report*'s conclusion on this point and reports on a study

[2] *See generally* Roy Stuckey et al., *Best Practices for Legal Education: A Vision and a Road Map* 235–273 (Clin. Leg. Educ. Assn. 2007); Karen Sloan, *Holding Schools Accountable: ABA Is Pushing Educators to Prove Their Law Graduates Can Cut It*, Natl. L.J. (Feb. 22, 2010) (available on LEXIS, Leg. News, Natl. L.J.) (discussing the ABA's proposed student learning outcomes accreditation standard).

[3] William M. Sullivan et al., *Educating Lawyers: Preparation for the Profession of Law* (Jossey-Bass 2007) [hereinafter *Carnegie Report*].

[4] *Id.* at 185–197.

[5] *Id.* at 186.

[6] *Id.* (defining the capacity to "think like a lawyer" as "capacities for understanding legal processes, for seeing both sides of legal arguments, for sifting through facts and precedents in search of the more plausible account, for using precise language, and for understanding the applications and conflicts of legal rules").

whose findings are relevant to an assessment of legal educators' effectiveness in teaching legal analysis. It is useful to begin the inquiry in this area by examining the research methodology used by the *Carnegie Report* research team in reaching their conclusion. The body of evidence that forms the basis for the researchers' conclusion that law schools are effective in teaching legal analysis consists of direct observation of classes at sixteen law schools, student and faculty interview responses, and survey data provided by the Law School Survey of Student Engagement (LSSSE) report.[7] Interpreting this evidence, the researchers conclude that their findings constitute "unmistakable evidence of the pedagogical power of the first phase of legal education,"[8] with this first phase (i.e., the development of skills of legal analysis) being accomplished early in the course of legal education—certainly within the first year, probably beginning early in the first semester.[9]

The first section of this Article raises questions about the *Carnegie Report*'s conclusions concerning the effectiveness of traditional legal education through a review of the empirical studies on law students' development of basic skills of legal reasoning. The second section of this Article reports the results of a pre- and post-test study that we conducted. We designed the study to assess student learning outcomes in a traditional first-year law school course that has an express goal of developing students' basic skills of legal reasoning. Consistent with the findings of prior studies,[10] the findings from our study reveal that the class as a whole did not evidence any significant positive movement in the development of legal reasoning skills. While some individual

[7] LSSSE is an extensive survey of law students concerning their experiences in law school, both in class and outside the classroom. *Id.* at 75–77.

[8] *Id.* at 186. While the researchers find that legal educators are effective in teaching skills of legal analysis, they fault legal educators for failing to move on to develop students' practical skills or professional values. For the authors, these missing components define the shortcomings of legal education today.

[9] *Id.* ("Within months of their arrival in law school, students demonstrate new capacities. . . .").

[10] *See* Kevin D. Ashley, *Teaching a Process Model of Legal Argument with Hypotheticals*, 17 Artificial Intell. & L. 321 (2009); David P. Bryden, *What Do Law Students Learn? A Pilot Study*, 34 J. Leg. Educ. 479 (1984); Dorothy H. Evensen et al., *Developing an Assessment of First-Year Law Students' Critical Case Reading and Reasoning Ability: Phase 2* (Law School Admissions Council Grants Report 08–02, Mar. 2008) (available at http://www.lsac.org/lsacresources/research/gr/pdf/gr-08-02.pdf).

students demonstrated learning gains as a result of their classroom experiences, such gains were not evenly or widely shared by the group as a whole. The final section of this Article discusses the implications for legal education. The findings of the new study, along with those from the prior studies, indicate that while current legal education may not diminish law students' reasoning skills, it may fail to enhance these skills. This Article concludes that the results across studies call for the continued development of teaching approaches in doctrinal courses other than the traditional case-dialogue method, along with the rigorous assessment of learning outcomes.

I. EMPIRICAL STUDIES OF THE DEVELOPMENT OF LEGAL READING AND REASONING SKILLS

A developing line of empirical research indicates that the effectiveness of current legal instruction in developing basic skills of legal analysis is an open empirical question.[11] While the *Carnegie Report* indicates that the participants in legal education perceive a remarkable intellectual transformation,[12] their perception of learning gains needs to be tested. Such testing will help determine whether the perception is grounded in fact, and if it is, when the learning gains occur, how they are achieved, and how legal educators can increase these gains.

A. The Particular Skill of Legal Analysis Examined

The previous studies in this area indicate that there are multiple discrete reasoning tasks that constitute basic skills of legal analysis.[13] This Article examines a particular reasoning task that each of the prior studies addresses at least to some degree, "cross-case/hypothetical reasoning." One research team led by education psychologist Dorothy Evensen and communications researcher James Stratman provides a useful initial description of this particular aspect of legal analysis:

[11] *See* Ashley*, supra* n. 10; Bryden, *supra* n. 10; Evensen et al., *supra* n. 10.

[12] *Carnegie Report*, *supra* n. 3, at 47, 186.

[13] *See* Ashley, *supra* n. 10, at 330–334; Bryden, *supra* n. 10, at 480–481; Evensen et al., *supra* n. 10, at 2–5.

> [Law] students need to be able (a) to construct accurate representations of multiple, closely related cases, (b) to detect indeterminacies of interpretation arising between or among them, and (c) to distinguish more from less purpose-relevant questions about their relationships to each other.[14]

The research team describes this particular skill of legal analysis as cross-case reasoning that addresses indeterminacies in text and meaning. The team labels test questions designed to assess this skill as "cross-case, indeterminate" items.

Another research team led by Kevin Ashley focuses on a closely related form of legal analysis—"hypothetical reasoning."[15] Hypotheticals are, in essence, "novel cases presenting new dilemmas."[16] These "novel cases" are used, at least in part, to create "conceptual bridges between cases along a continuum."[17] (This function of hypothetical reasoning parallels the construction of accurate representations of multiple, closely related cases as described by the Evensen/Stratman team.) Hypothetical reasoning also involves drawing analogies and distinctions between or among case situations in the context of conflicting principles and policies. They allow one to "explore a space of situations that may or may not be distinguished on a normative and policy basis from the case at hand,"[18] with indeterminate principles and policies informing one's judgment of which similarities and distinctions are relevant. (In short, this function of hypothetical reasoning parallels the detection of indeterminacies among cases and the judgment to identify relevant questions about the relationship among cases as described by the Evensen/Stratman team.) Because of the functions they serve, hypotheticals play a prominent role in legal education, inducing "the student to apply the concept under discussion to other patterns of facts, thereby revealing the

[14] Evensen et al., *supra* n. 10, at 4. Indeterminacies of interpretation are a function of linguistic (semantic and syntactic) aspects of a single text or arise between/among multiple texts (ambiguities, contradictions, silences, vagueness, etc.). Accordingly, one cannot conceptualize indeterminacies in absolute terms; they are matters of degree. *Id.* at 3–4.

[15] Ashley, *supra* n. 10, at 331.

[16] *Id.*

[17] *Id.* at 331 n. 6 (quoting Robert S. Summers & Sven Eng, *Departures from Precedent*, in *Interpreting Precedents: A Comparative Study* 528–529 (Ashgate Publishers 1997)).

[18] *Id.* at 331.

student's grasp of the underlying legal notion, and to derive, explore, and refine legal principles."[19]

So described, cross-case reasoning based on textual indeterminacies and hypothetical reasoning are very similar. These similar forms of reasoning constitute the aspect of legal analysis that lies at the core of the inquiry addressed in this Article—cross-case/hypothetical reasoning. The focus on this aspect of legal analysis provides the context for the discussion of prior studies in this initial section of the paper.

B. The Bryden Study

To determine the feasibility of assessing selected analytical skills, David Bryden conducted a pilot study in the area of legal reasoning skills.[20] His study, published in 1984, found that third-year law students demonstrated legal reasoning skills superior to those of entering law students. However, he also found that the legal reasoning skills of third-year law students were remarkably weak. Thus, Bryden's findings called into question the effectiveness of traditional legal education in teaching skills of legal analysis.[21]

One skill that Bryden termed "functional analysis," seemingly because it focused on the intended function or purpose of a legal rule, required students to engage in cross-case reasoning. Bryden explained that, for purposes of his study, functional analysis calls on students to identify and draw on the purpose of a legal rule or category as defined by a set of opinions and/or statutory provisions in order to complete a lawyering task.[22] The study was designed to test law students' ability to recognize a situation where a functional analysis may be appropriate. Bryden's hypothesis was that a good legal education would produce students who at least recognize and articulate the possibility of resolving a relevant legal ambiguity by reference to the purpose of a legal rule or category.

To test this hypothesis, Bryden developed two examinations, each of which consisted of essay questions that presented students with a short set of facts that constituted a legal problem.

[19] *Id.* at 333 (citing and quoting *Carnegie Report*, *supra* n. 3, at 62, 66, 68, 75).

[20] Bryden, *supra* n. 10.

[21] *Id.* at 500–503.

[22] *Id.* at 481, 485–491.

The questions also provided students with a set of short hypothetical judicial opinions and/or statutory provisions. The questions asked students to perform a specific legal task—write a memorandum analyzing the legal problem. To diminish the effect of prior substantive legal knowledge and to focus on students' capacity to engage in cross-case functional analysis, the examinations provided participants with all the relevant legal materials necessary to address the legal problem.[23]

At three law schools, Bryden invited two separate groups of students to complete the two tests. The first group consisted of third-year law students enrolled in their last semester who accepted an invitation to complete one of the two tests. The second group consisted of incoming students at the same three law schools. The incoming students were invited to participate based on their LSAT scores, which the researchers had selected in order to match, as much as possible, those of the students in the first group. This second group of students took the tests during the summer immediately after the first group had completed law school and immediately before they began law school.[24]

Overall, the results of the study indicated that third-year law students were "nearly always more proficient" in terms of functional analysis than the entering students.[25] Thus, the study indicated that some incremental learning gains in this skill resulted from three years of legal education. However, Bryden noted that even for the third-year students, a clear majority of the essay answers failed to indicate any engagement in functional analysis whatsoever.[26]

Based on the study's overall results, Bryden concluded,

> Every [third-year student] who took one of the tests had undoubtedly been taught—not once, but several times—[the skill of functional analysis that was] necessary to make the types of points that the questions were designed to elicit. But in nearly every instance, most of them did not even see

[23] *Id.* at 481, 484–485.

[24] *Id.* at 482–483.

[25] *Id.* at 500.

[26] *Id.*

> the issue. The moral seems to be that what we "teach" is not what they learn.[27]

Thus, the study's results failed to support Bryden's hypothesis that what is perceived as a good legal education produces students who at least recognize and articulate the possibility of resolving a relevant legal ambiguity by engaging in functional analysis. Bryden's findings raise serious questions about the effectiveness of traditional legal education in teaching the skill of functional analysis.

Education researcher James Stratman has provided a wide-ranging critique of Bryden's study.[28] Stratman questioned the use of essay questions to assess skills of legal analysis such as functional analysis. He noted that students may have engaged in appropriate analytical activity, but failed to include evidence of their analytical process in their writing.[29] Bryden provided support for Stratman's concern when he noted that some students reached a sensible conclusion to the legal problem while failing to advance a thoughtful reason for the conclusion.[30] In the end, Stratman asserted that Bryden's study is not so much valuable for its results (third-year students perform better than entering students), but rather for its support of an effort to develop empirical tests that can more effectively examine and assess the various skills of legal analysis.[31]

C. The Evensen/Stratman Study: Phase 1

About two decades after Bryden published his study, another research team that included Stratman and was led by Dorothy Evensen took up the effort to develop empirical tests.[32] This team published the results of its study in 2008. While the development of valid test instruments was the primary goal of their study, the Evensen/Stratman team found that the legal reasoning skills of second-semester law students were no better than those of first-semester law students.

[27] *Id.* at 503.

[28] James F. Stratman, *The Emergence of Legal Composition as a Field of Inquiry: Evaluating the Prospects*, 60 Rev. Educ. Res. 153, 166–171 (1990).

[29] *Id.* at 168.

[30] Bryden, *supra* n. 10, at 502.

[31] Stratman, *supra* n. 28, at 171.

[32] Evensen et al., *supra* n. 10.

In designing their study, the Evensen/Stratman researchers identified the study's practical purposes: "to develop a prototype multiple-choice instrument assessing law students' critical case reading and reasoning skills."[33] They also acknowledged that this area of inquiry is under-researched despite its importance to legal educators trying to instruct students in the basic reasoning skills required for legal discourse. Based on the success of quantitatively scored tests in identifying underlying constructs in comparably challenging non-legal analysis tasks (e.g., SAT, LSAT, and some multiple-choice reading comprehension tests), they asserted that well-designed multiple-choice tests have the potential to address this gap in the research.[34]

The researchers posited that a valid test of case reading and reasoning skill should assess students' ability "(a) to construct accurate representations of multiple, closely related cases, (b) to detect indeterminacies of interpretation arising between or among them, (c) to distinguish more from less purpose-relevant questions about their relationships to each other."[35] Such an assessment allows the identification and measurement of the difficulties students face in cross-case reasoning, difficulties that are likely distinct from those encountered in single-case reasoning (i.e., the ability to identify accurate paraphrases of determinant meanings and the ability to read and analyze indeterminacies of interpretation within a single-case text).[36]

Based on these testing goals, the researchers designed a test instrument that required students to read three cases that address the same procedural rule. There were significant indeterminacies of interpretation both within each case and across cases. The test-taker was given a specific purpose for reading and thinking about the cases—to prepare an appeal of a decision against the defendant in one of the three cases.[37]

[33] *Id.* at 1.

[34] *Id.* at 2.

[35] *Id.* at 4 (In defining "indeterminancies," the researchers stated that they were "speaking of all the rhetorical ways that statements offered by courts may be open to interpretation, such that there may be no way to tell precisely what a court means or precisely how it is reaching its conclusion.").

[36] *Id.* at 7–8.

[37] *Id.* at 4–5.

The test consisted of fourteen questions, with each question falling into one of four categories:[38]

1. Four questions were designed to test students' ability to read a single case and accurately recognize case content regardless of their purpose in reading the case ("single-case, determinate" questions). (For example, one test item asked students to select the statement that best summarizes the dissent's reasoning in a single case.)

2. Four questions tested students' ability to read multiple cases and accurately identify case content regardless of their purpose in reading the cases ("cross-case, determinate" questions). (For example, one test item asked students to select the response that best summarizes the legal issue raised by all three cases.)

3. Three questions tested students' ability to read a single case and to identify and/or evaluate indeterminacies of interpretation relevant to their purpose in reading the case ("single-case, indeterminate" questions). (For example, one test item asked students to take on the role of an attorney appealing a decision on a particular issue, to consider specific ambiguities presented by a case they have been asked to read, and to identify the ambiguity that is most relevant to their appeal.)

4. Three questions tested students' ability to read multiple cases and to identify indeterminacies of interpretation across cases that are relevant to their purpose for reading the cases ("cross-case, indeterminate" questions). (For example, one test item asked students to determine the most important question raised by two prior appellate court decisions in light of their assignment to construct a theory of the case for an appeal of a trial court decision.)

The researchers' field test of the test instrument was conducted at five law schools and involved 161 first-year law students from five law schools who volunteered to participate, 81 of whom took the test in their first semester (fall 2003) and 80 of

[38] *Id.* at 5, 50–58.

whom took the test in their second semester (spring 2004). The students who completed the test in the fall were not the same students who completed the test in the spring, but the students were matched across the two groups based on LSAT scores. Thus, this phase of the study (phase 1) used a matched-subjects design.[39]

Prior to test administration, the researchers had formulated a hypothesis that predicted "that the mean score of the students taking the test in their second semester would be significantly higher than that of students taking the test in their first semester."[40] The results of the study failed to support this hypothesis concerning student improvement or development. The fall and spring students' mean test scores were not significantly different despite matching mean LSAT scores. (The fall students' mean score was 7.96 correct, while the spring students' mean score was 7.86 correct.[41]) This lack of significant difference also held true when each of the four question types was examined separately, indicating no statistically significant improvement for any particular skill category, such as cross-case, indeterminate reasoning.[42]

[39] *Id.* at 4, 6.

[40] *Id.* at 6.

[41] *Id.*

[42] *Id.* at 6–7. The researchers had formulated two additional hypotheses. The first predicted "that the cross-case indeterminate test items would prove most difficult, while single-case determinate questions would prove least difficult." *Id.* at 7. The results provided support for this hypothesis concerning the relative difficulty of the test item types. Students provided the most correct answers for single-case, determinate questions, followed in order by cross-case, determinate; single-case, indeterminate; and cross-case, indeterminate. *Id.* Thus, the most difficult question type was cross-case, indeterminate. (Interestingly, the researchers found a very low correlation between performance on single-case questions and cross-case questions. The researchers also found a low correlation between performance on determinate questions and indeterminate questions. Thus, the test items appear to measure separate and independent reasoning skills. *Id.* at 7–8.)

The second hypothesis predicted "that students' overall scores on [the study's] test would correlate positively with their LSAT scores and also with their first-year [law school] GPAs." *Id.* at 8. The results generally supported this hypothesis concerning a correlation with other measures of achievement. Although the correlations were relatively weak, there were positive correlations with both LSAT scores and first-year GPAs. *Id.* These positive correlations provided some support for the external validity of the test, while the relative weakness of the positive correlations suggested that the test called for skills that were somewhat distinct from those assessed by the LSAT and law school grades. *Id.* at 9, 24.

D. The Evensen/Stratman Study: Phase 2

Despite the indication of external validity provided by the initial study and the use of a panel of legal experts to achieve test-construct validity, the researchers recognized that they needed to embark on a second study that would determine whether the test design and results could be replicated.[43] While this second phase of the study had a primary goal of developing valid test instruments, the Evensen/Stratman team found that law students' legal reasoning skills did not improve from the second semester to the third semester. In addition, they found that law students' legal reasoning skills did not improve between their first and third years of law school.

Phase 2 of the study involved the development of a second test that the researchers intended to be parallel, or equivalent, to the original test.[44] The researchers recognized that the development of this second test would allow them to address a methodological weakness of phase 1 of the study related to the finding that students' case reading and reasoning ability did not improve from the first semester to the second semester. This finding was subject to doubt because the researchers had used a matched-subjects design, with matching based only on LSAT scores. In other words, phase 1 of the study had not tested the same students twice, once during the first semester and once during the second semester. This was because the researchers had only developed one test. The development of a second test would allow the researchers to use a superior within-subjects design, testing the same students at two different points in time.[45]

In the phase 2 field test, the researchers posed two specific research questions that are relevant to the discussion here. First, "Do students' case reading and reasoning skills improve between their first and second years in law school?" Second, "Do students' case reading and reasoning skills improve between their first and third years in law school?"[46]

For the phase 2 field test, the researchers initially recruited 146 first-year students from the same five law schools included in

[43] *Id.* at 9.

[44] *Id.*

[45] *Id.*

[46] *Id.* at 14.

phase 1. The researchers randomly assigned these students to take either the original version of the test or the new version of the test in the spring 2006 semester. Eighty-three of these students completed the full study protocol, taking the alternate version of the test in the fall semester of their second year (fall 2006). Forty-nine of these students took the original version of the test first, and thirty-four took the new version of the test first. This procedure allowed for a within-subjects comparison of reading and reasoning skills, as revealed by test performance, from the second semester of the first year to the first semester of the second year.[47]

The phase 2 field test also involved sixty-three third-year law students who had participated in phase 1, and thus had completed the original test in their first year, either in the fall or the spring semester. These students completed the second test in their last semester of law school (spring 2006). This allowed for a within-subjects comparison of reading and reasoning skills from the first year to the third year.[48]

Consistent with the results from phase 1, the researchers found that students' case reading and reasoning skills did not improve between their second and third semesters.[49] In addition, the researchers found students' skills, as measured by the tests developed by the researchers to date, did not improve between their first and third years of law school.[50]

For purposes of the current study, it is important to note that phase 1 and phase 2 of the Evensen/Stratman study indicated that cross-case, indeterminate reasoning is the most difficult skill for law students to demonstrate.[51] In addition, the Evensen/Stratman study indicated that law students do not improve their cross-case, indeterminate reasoning skill as they progress in their legal education.[52] These results call for both additional edu-

[47] *Id.* at 14–15.

[48] *Id.* at 16.

[49] *Id.* at 15, 27.

[50] *Id.* at 16, 27. As in phase 1, the results in phase 2 showed a decrease in performance from single-case type questions to cross-case type questions and from determinate type questions to indeterminate type questions. *Id.* at 16–19, 27. Once again, single-case, determinate items were the easiest for students and cross-case, indeterminate items were the most difficult. *Id.* at 16–19. In addition, the researchers found very similar positive correlations with LSAT scores and law school GPAs for both test versions. *Id.* at 24–26.

[51] *Id.* at 7, 16–19.

[52] *Id.* at 6–7, 15–16, 27.

cational interventions designed to improve this skill and additional assessments of this skill. They also may call for discussion among legal educators of the importance of this skill to law school curricula and precisely when and where this skill should be the direct or indirect subject of instruction.

E. The Ashley Studies

In 2007, in contrast to Evensen and her colleagues, Kevin Ashley conducted a set of studies that addressed whether an educational intervention beyond traditional law school classroom instruction improved legal reasoning skills.[53] While the primary purpose of his studies was to assess the educational outcomes of his intervention, his studies also allowed for an assessment of traditional legal education because they involved students who were concurrently enrolled in traditional first-year courses. In the largest study, the law students failed to achieve any significant gains in legal reasoning skills as a result of their participation in either traditional law school classes or the educational intervention introduced by Ashley.

As to the primary focus of Ashley's studies, the educational intervention was designed to help students learn to reason with hypotheticals in the context of personal jurisdiction doctrine through an intelligent tutoring system. The system helped students identify, analyze, and reflect on episodes of reasoning with hypotheticals in oral argument transcripts through the construction of simple diagrams.[54]

Ashley's hypothesis was that law students who used the program to diagram hypothetical reasoning would learn hypothetical reasoning skills better than law students who studied hypotheticals without the program's diagramming support and feedback.[55] Ashley's studies tested his hypothesis through the construction and application of pre- and post-test instruments. The pre-test and the post-test both used three multiple-choice question types to measure learning gains—items that assessed argument skills in contexts that require no background legal knowledge; items that explored the use of hypotheticals in making arguments in a

[53] Ashley, *supra* n. 10, at 332–336.

[54] *Id.* at 322, 339.

[55] *Id.* at 347.

non-legal, intuitive domain; and items based on the conceptual model of hypothetical reasoning that explored judges' use of hypotheticals and advocates' engagement with hypotheticals. The post-test included two additional multiple-choice question types used by the researchers to compare an experimental group with a control group—"[n]ear-transfer legal argumentation questions involving the selection of proposed legal tests, hypotheticals, and responses in a personal jurisdiction case that was new to the students"[56] and "far-transfer legal argumentation . . . questions involving hypothetical reasoning similar to that in the personal jurisdiction cases but drawn from a new legal domain (copyright law) with which first-year students were not likely to be familiar.[57] Most of the test questions engaged students in reading and reasoning activities that are similar to those called for by the Evensen/Stratman study's cross-case, indeterminate questions.[58]

In the largest of his studies, Ashley required participation by all eighty-five students in one section of the fall 2007 Legal Process course at one law school. This course covered personal jurisdiction doctrine, the area of law addressed by the oral arguments used in the study.[59] The students were randomly assigned to one of two study conditions, balanced by LSAT scores. The experimental group received training through a graphical tutoring program that supported argument diagram creation and gave advice. The control group completed a text-based training program that used the same oral arguments as the graphical tutoring program but did not provide feedback.[60] For both groups, the period between the pre-test and post-test was five weeks. The specific oral arguments and cases used on the tests were not covered in class discussions.

[56] *Id.* at 349.

[57] *Id.* at 349–350.

[58] Notably, Ashley's test items "were not formally checked for validity and reliability. However, they have face validity as reported by an experienced law school professor and some advanced graduate law students who took the test." Neils Pinkwart et al., *Evaluating an Intelligent Tutoring System for Making Legal Arguments with Hypotheticals*, 18 Intl. J. Artificial Intell. Educ. 401, 410 (2009). In addition, several of the test items were modeled on questions drafted for possible inclusion in the LSAT. *Id.*

[59] Ashley, *supra* n. 10, at 348, 353. The researchers used oral arguments presented before the United States Supreme Court in the following personal jurisdiction cases: *Burnham v. Super. Ct. of Cal., Co. of Marin*, 495 U.S. 604 (1990); *Asahi Metal Industry Co. v. Superior Court of California*, 480 U.S. 102 (1987); *Burger King Corp. v. Rudzewicz*, 471 U.S. 462 (1985); and *Kathy Keeton v. Hustler Mag., Inc.*, 465 U.S. 770 (1984).

[60] Ashley, *supra* n. 10, at 348.

The results revealed no statistically significant differences between the experimental and control groups with respect to performance on the post-test.[61] The results also indicated that neither group benefitted from the study in terms of improvement from pre-test performance.[62] The only significant difference between the pre-test and post-test scores was a decline on one question type for the students in the experimental group with comparatively low LSAT scores.[63] Thus, the combination of traditional law school classroom case discussion of personal jurisdiction cases and either the diagramming program or the text-based program resulted in no significant educational gains in terms of hypothetical reasoning skill from pre-test to post-test.[64]

II. THE CURRENT STUDY

The study reported in this Article is the first step in a larger project that seeks to extend the Evensen/Stratman and Ashley studies by employing multiple-choice pre- and post-test instru-

[61] *Id.* at 350, 352–353.

[62] *Id.* at 350.

[63] *Id.* at 353. (The test scores for students in the experimental group with an LSAT below the median of 159 dropped from pre-test to post-test with respect to the general questions on personal jurisdiction.).

[64] The Authors of this Article conducted a validation study of the tests used by Ashley. (The results of the validation study are on file with the Authors.) In this study the tests were provided to a panel of ten law school faculty at the University of Pittsburgh School of Law. These faculty members took the tests independently of the training and thus received them without any additional instruction, simply reflecting on their pre-test experience and then taking the post-test. The faculty members were then compared to the first-year students who participated in Ashley's large study and a group of twenty-three third-year students who volunteered to use the system as part of a separate study completed by Ashley. All three groups were subject to a statistical comparison in order to assess their relative performances. A preliminary analysis of the results indicated that the full set of third-year students significantly outperformed the first-year students on both the pre- and post-tests. However they did not achieve higher learning gains than the first-year students. This was also true for comparisons between the faculty and the first-year population. Interestingly, no significant difference was found between the faculty and the third-year student volunteers in terms of pre- and post-test scores or learning gains.

The performance differences between the third-year and first-year students support the contention that the skills being assessed are taught in law school or at least favor individuals with superior qualifications. These dual explanations are also supported by the differences between the faculty and first-year students. However, distinguishing between these explanations is complicated by the fact that both the faculty and third-year students were self-selected populations with the third-year students intentionally drawn from the upper-half of the law school class. Thus, while the results support the validity of the test as a measure of expertise and performance, further analysis is required to assess their indications for legal education.

ments to measure learning gains for first-year law students in the area of cross-case/hypothetical reasoning skills. This pilot study is more focused than the previous studies in terms of course education goals, teaching methodologies, subject matter coverage, and timeframe. Namely, the study is conducted within the context of a single first-year, first-semester course that has an express goal of improving students' skills of legal analysis, with an emphasis on the skill of cross-case/hypothetical reasoning. The faculty member teaching this course uses no educational intervention other than traditional classroom case-dialogue method. The study is centered on a single substantive law unit of the course—personal jurisdiction doctrine. The course completes this unit of study during the first six weeks of the semester. Thus, the study focuses on this discrete period of instruction, with students completing the pre-test in week #1 of their law school education and the post-test in week #7.

The study utilizes pre- and post-tests derived from the multiple-choice tests developed by Ashley. The majority of the questions were designed to require students to demonstrate the skill of cross-case/hypothetical reasoning. The study provides a starting point for future research on student learning outcomes in this important area of legal education.

This pilot study was conducted at the University of Pittsburgh School of Law. As described below, the School provides a traditional first-year curriculum. It enrolls a first-year class of approximately 240 students. In terms of LSAT scores and undergraduate GPAs, student qualifications are typical for law schools that rank in the second tier of the U.S. News and World Report law school rankings.[65]

First-year students at the School are divided into three sections. Every section is required to take the same set of substantive law courses. Students in each section have the same instructors for their doctrinal/substantive law courses. Students are randomly assigned to sections. For the purposes of this study, we tested students in one section of the fall 2009 first-year class in

[65] U.S. News & World Rpt., *Best Law Schools for 2010*, http://grad-schools.usnews.rankings_andreviews.com/best-graduate-schools/top-law-schools (accessed Sept. 28, 2012). For the relevant year (2009–2010), the second tier consisted of schools that ranked 52 to 98, with three schools tied at 98. *Id.* The University of Pittsburgh ranked 67 that year, with an LSAT 25th/75th percentile range of 157–161 and an undergraduate GPA 25th/75th range of 3.18 to 3.63. *Id.*

one class—Legal Process (i.e., Civil Procedure I).[66] One instructor taught this section of the course—an experienced, competent legal educator who utilizes the traditional case-dialogue method of instruction. The total section size was seventy-seven.

The students in the section were divided into two groups using balanced, random assignment by maximum LSAT score.[67] This assignment was designed to ensure that the subsequent comparison would not be affected by any differences in terms of an established measure of incoming competence between the groups. In addition to the maximum LSAT score received, we also obtained other demographic information about the students, including their gender, ethnicity, and undergraduate GPA. Ultimately, due to absences and attrition, seventy-one students, thirty-seven in group 1 and thirty-four in group 2, completed the entire study. We base our subsequent analyses on this set.

Each group followed the same study process. Students began by taking one of two tests designed to assess their reading and reasoning abilities as the pre-test. They then received six weeks of in-class instruction in the subject area of personal jurisdiction, followed by the alternative test as the post-test. As described in more detail below, each test asked students to answer similar analytical questions in the context of one of two possible personal jurisdiction cases. Apart from the order of the tests (as explained in the Results section below), the two groups received the same instructional experience.

[66] The other substantive law courses that semester consisted of Contracts, Criminal Law, and Torts. The students were also required to take a Legal Analysis and Writing course. For this course, the section was divided into three small sections with approximately twenty-five students in each. A single instructor taught two of the small sections, with a second instructor teaching the remaining small section. For purposes of the study, students were assigned to the two study groups so that each small section had equal representation in each study group. The findings did not indicate any significant difference in test performance or learning gains by small section/legal writing instructor assignment. During the relevant six-week study period, students in the legal writing course completed introductory exercises on legal analysis that included understanding a case text, applying the law to facts, discussing a short judicial opinion in class, and analyzing a hypothetical in class. Students often worked in small groups in completing these exercises. The students also completed a closed memo assignment that involved several short judicial texts. They received written feedback from their instructor before the end of the six-week period.

[67] As students are permitted to take the LSAT multiple times, we elected to use their maximum score reported to the law school as a basis of comparison. The resulting groups (thirty-nine in group 1 and thirty-eight in group 2) showed no significant difference in terms of the maximum LSAT scores reported (one student in group 2 did not have an LSAT score in his or her student file).

For the purposes of this study we constructed two multiple-choice tests centered on an oral argument to the United States Supreme Court in a personal jurisdiction case. One test was based on an argument in *Keeton v. Hustler Magazine*,[68] while the other was constructed around an argument in *Calder v. Jones*.[69] Both tests were designed by legal experts and designed to be of equal difficulty. One group of students took the *Keeton* test as its pre-test and the *Calder* test as its post-test, while the other group of students did the opposite.

Students taking the tests first read some short background about the case followed by an extract of the pertinent oral argument. They then answered questions designed to test their reading and reasoning skills. Each test consisted of ten questions, seven of which were cross-case, indeterminate questions as described above and three of which were single-case, determinate questions.

A sample cross-case, indeterminate question from the *Keeton* test is shown below (with a hypothetical constituting one of the cases in the cross-case comparison). Here the student is required to assess a hypothetical that was raised by a Justice during the oral argument in light of a legal test proposed by the advocate:

Keeton—Which Hypo Is Problematic?

Assume that [the test proposed by Mr. Grutman, the petitioner's attorney] 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

[68] 465 U.S. 770 (1984),

[69] 465 U.S. 783 (1984).

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.[70]

By contrast, a single-case determinate question from the *Calder* test is shown below. In this question the student is asked to identify the central legal issue presented in the argument made in *Calder v. Jones*:

Calder—Legal Issue

What legal issue concerning Calder and South's appeal in Shirley Jones v. Iain Calder and John South did the Justices

[70] It should be noted that this test item is presented only for the purpose of illustrating the question type and format. It is not presented in order for the reader to determine the correct responses. (Selections a. and d. are considered correct.) Interested readers can obtain the full test instrument from the Authors upon request.

address in the oral argument excerpt? Select the best answer below.

a. In determining whether the courts in a state may exercise personal jurisdiction over a defendant, should the Court consider the defendant's travel to the forum state that was unrelated to the events giving rise to the suit?

b. May courts in a state exercise personal jurisdiction over an out-of-state individual who actively participated in the investigation and production of a nationally distributed magazine story about an in-state plaintiff?

c. Should an employee be entitled to heightened protection from personal jurisdiction when his or her employer is a defendant, concedes jurisdiction in the forum, and has the capacity to pay a damage award?

d. Should a defendant in a libel suit be immune from personal jurisdiction in a distant forum because of First Amendment free speech concerns?[71]

Group 1 began with the *Keeton* test while Group 2 began with the *Calder* test. Both groups received the same in-class instruction from the same instructor before taking the post-tests, with Group 1 taking the *Calder* test as a post-test and Group 2 taking the *Keeton* test. The instructor did not ask students to read the *Calder* and *Keeton* cases and did not discuss either case in class. In the next section we will summarize the results of our student analyses and the performance both within and across groups.

III. RESULTS

In analyzing the results from this study we initially focused on addressing questions of test validity, that is, were the tests valid and equivalent measures of student performance? We also focused on overall performance, that is, did the students improve in terms of legal reading and reasoning skills? In addition, we sought to determine if the observed student performance is specif-

[71] It should be noted that this test item is presented only for the purpose of illustrating the question type and format. It is not presented in order for the reader to determine the correct response. (Selection b. is considered correct.) Interested readers can obtain the full test instrument from the Authors upon request.

ic to given subsets of the population where appropriate. For example, do students of one gender perform better than the other? While we did not deliberately balance the groups by gender, the random assignment gave us a representative split as we discuss below. And finally, we examined whether student performance varied by type of question (i.e., single-case, determinate; cross-case, indeterminate) and whether student performance on the study's tests correlated with student performance on the final examination for the entire Legal Process course.

Five of the ten multiple-choice questions on each exam were single-answer questions of the type shown in the *Calder* example above. That is, they asked students to select the only correct answer or the best answer. The remaining five questions on each test were multi-answer questions of the type shown in the *Keeton* example above where students were asked to pick all plausible answers.

When grading the exams we applied two distinct grading metrics: basic and even. Under the basic-grading metric students received points for correct answers only. On single-answer questions students received one point for each correct answer and zero points for each incorrect answer. In the *Calder* question shown above the student would receive one point for selecting the correct answer and zero points for all other choices. For multi-answer questions, students received fractional credit for each correct choice and zero for incorrect choices. In the multi-answer *Keeton* question shown above there are two correct answers. Selecting either one would bring students one-half of a point while selecting either of the other two choices would bring zero points.

While the basic-grading metric can be used to assess overall performance, it is, or can be, problematic if the students adopt a broad selection strategy on the multi-choice questions and hedge their bets by selecting most of the available choices. With that in mind, we applied an even-grading metric to the tests as well. Under this metric students receive a fraction of a point for each correct choice and lose a similar fraction for each incorrect choice. Given, for example, a multi-answer question with two correct choices and three incorrect ones, the students would receive one-half of a point for each correct choice and lose one-third of a point for each incorrect choice. Selecting all possible choices thus results in a net score of zero points for the question.

When assessing students' learning gains we considered both their *raw learning gains* (raw gain) and their *normalized learning gain* (NLG). Raw gain scores were computed by subtracting a student's pre-test score from his or her post-test score. While this is a natural way to compute individual gains, it fails to distinguish between poor students who make small gains and good students for whom only small gains are possible. For example, in comparing raw gain scores, a student who received a pre-test score of one point and improved to three points on the post-test would receive a raw gain score of two points as would a student who went from eight points on the pretest to ten points, the maximum possible score, on the post-test. This would mistakenly view these two students as equivalent.

For this reason we computed the normalized learning gain as the following:

(Post-test score – Pre-test score) / (Max Score – Pre-test Score)

That is, we compute the score as a function of the raw gain divided by the maximum gain possible. Under this metric the student who moved from one to three points would receive a normalized learning gain score of 2/9 while the student who moved from eight to ten points would receive a score of 1 (2/2).

A. Validity Analysis

In the work of Evensen and Stratman the test questions were designed and tested by expert law-school faculty to ensure that they were valid measures of the students' performance.[72] As we discussed above, a similar route was taken in connection with the work by Ashley, with questions being designed and tested by legal experts as well as being subject to a subsequent faculty-student comparison.[73] Given this prior work and the fact that the existing tests drew on the work of Ashley, our primary goal in establishing test validity was to determine whether the *Keeton* and *Calder* tests were of equal difficulty. If the two tests are equally difficult then the order in which the students receive them should not matter. If, however, the tests do differ in difficulty then students

[72] Evensen et al., *supra* n. 10, at 9, 13, 21.

[73] *See supra* nn. 58, 64.

who receive the easier test first would likely show lower learning gains than comparable peers.

As stated above we randomly assigned students to conditions by their LSAT scores. Group 1 had a mean LSAT score of 159.41 (sd=3.86[74]) while group 2 (excluding the single student who did not report a score) had a mean score of 159.88 (sd=3.57). A t-test[75] comparison of the LSAT means reported no significant difference in LSAT scores between the two groups (t(37)=-0.54, p>0.59, power = 0.49[76]).

[74] "sd" represents the standard deviation of a population score or random variable. It is a measure of the spread of scores about the mean. Higher standard deviations indicate that the scores are more widely distributed than lower standard deviations. *See* William L. Hays, *Statistics* (Harcourt Brace College Publishers 1994).

[75] A t-test is a statistical test designed to test whether the mean score of a population (e.g., pre-test scores) differs between two groups (*two-sample* t-test) or differs significantly from a baseline value such as 0 (*one-sample*). The t-tests presented above are of the latter type. T-tests work by testing the null hypothesis that the populations match, or the single population is the same as the baseline, and rejecting it if the probability of obtaining the sample data given the null hypothesis lies below a set threshold (conventionally 0.05, which is the threshold used in this study). T-tests are typically used to test a *two-sided* hypothesis—do the two populations match or differ? T-tests can also be used in a *one-sided* fashion to test whether one population mean is greater of lesser than the other or whether the single population mean exceeds or falls below the baseline. The tests reported above are *one-sample, one-sided* tests.

The t-test produces a t-statistic from which we compute the p-value, which is the probability of obtaining the observed result if the null hypothesis is correct and thus falsely rejecting the null hypothesis. In statistics, this is known as making a *type-I* error. If the p-value is below a set threshold then we can say with confidence that future tests of the same hypothesis will also find a difference. *See* Peter Dalgaard, *Introductory Statistics with R* (Springer Verlag 2002).

[76] The "power" of a t-test is one minus the probability of making a *type-II* error. In other words, it is the probability that the null hypothesis is false (i.e., the two groups differ), but that we did not detect the difference. Given a t-test that reported no statistical difference (by convention, power is not reported if the p-value is below the chosen threshold), we can also view the power as the probability that the groups do not differ. For results with high power, we can say with confidence that the two groups are the same and that future tests will also find no difference. For tests with low power we cannot make such guarantees. In practical terms, obtaining high power for a test requires a larger number of students than obtaining low p-values. In other words, robust tests for rejection of the null hypothesis are easier to fulfill than robust tests for a non-difference. *See* Dalgaard, *supra* n. 75.

Here the test had a statistical power of 0.51, indicating that there exists a 49% chance that no true difference exists. While this power is relatively low, the absence of any significant differences between the two tests leads us to conclude that the tests are sufficiently balanced for the present purposes and, indeed, the test items have been more carefully analyzed and balanced than most law school writing assignments and exams. A larger scale study would be required to form a more complete balance. We have been considering this as future work.

Under the basic-grading rubric, there were no significant differences between the groups using a basic analysis of variance.[77] There were also no significant differences between the groups under the even-grading rubric.[78] These results, coupled with the similar structure of the within-group differences for each group, lead us to conclude that the order of the tests was immaterial and that, by extension, the tests are equivalent.

B. Overall Performance

Having shown that the tests are equivalent, we turn to the student performance for the entire study population. These results are designed to address our primary question regarding student learning gains: Did the students improve overall from pre- to post-test?

Under the basic grading rubric, the mean pre- and post-test scores were: 4.68 (sd=1.7) and 5.14 (sd=1.53), respectively. Both scores were significantly greater than 0 (pre-test: $t(70)=23.19$, $p<6.9e\text{-}34$;[79] post-test: $t(70)=28.17$, $p<3.19e\text{-}39$). Comparing the individual raw gain scores from pre- to post-test we found that the mean gain was 0.46 (sd=2.21). A t-test confirmed that this score was significantly greater than the baseline 0 ($t(70)=1.75$, $p<0.04$). This was not, however the case for the normalized learning gain. Here the mean score was -0.02 with a standard deviation of 0.53. However, a two-sided t-test found that this score was not significantly different from the baseline of 0 ($t(70)=-0.45$, $p<0.67$, power = 0.02). This shift from a positive learning gain of 0.46 to -0.02, which indicates the apparent lack of sensitivity of

[77] Pre-test: $F(1,64.1)=0.56$, $p=0.45$; post-test: $F(1,67.4)=0.06$, $p=0.81$; raw gain: $F(1,63.7)=0.56$, $p=0.46$; and NLG: $F(1,59.1)=1.25$, $p=0.27$. Analysis of variance (anova) tests, like the t-test discussed *supra* n. 75, are designed to assess the differences in mean scores between groups. Unlike the t-test, these analyses generalize to more than two groups. The one-way anova tests applied here are also robust for groups with unequal variances. The statistical result of an anova test is an F-score that is conventionally reported in the form $F(n,d) = f$, where n is the number of degrees of freedom in the numerator and d is the number of degrees of freedom in the denominator. The f-statistic is drawn from the f-distribution and from this statistic we compute the reported p-value that, as with the t-test, shows the probability of receiving the observed result if the null hypothesis (i.e., agreement among all group means) holds. *See* Dalgaard, *supra* n. 75.

[78] Pre-test: $F(1,65.2)=0.58$, $p<0.45$; post-test: $F(1,68.7)=0.22$, $p<0.64$; raw gain: $F(1,65.1)=0.86$, $p<0.36$; and NLG: $F(1,57)=1.81$, $p<0.18$.

[79] Engineering notation is used to represent extreme values. In engineering notation, the e delimiter indicates the exponent part for an order of magnitude multiplication. Thus, 6.9e-34 is 6.9 * 10 to the 34th power.

the basic-grading rubric is, in part, what motivated our decision to apply the even-grading rubric.

Under the even-grading rubric, however, the pattern was similar. The pre- and post-test scores were significantly greater than 0 (pre-test: mean=3.79, sd=1.82, t(70)=17.51, p<1.37e-26; post-test: mean=4.43, sd=1.76, t(70)=21.18, p<1.83e-32). And again the raw gain scores were significantly positive, although less than a full point (mean=0.64, sd=2.4, t(70)=2.24, p<0.02), while the normalized learning gain was not significantly different from the baseline (mean=0.02, sd=0.44, t(70)=0.39, p<0.69, power = 0.05).

The table below reports the results on overall performance:

Basic-Grading Rubric			
Score	**Mean**	**sd**	**Comparison with baseline 0**
Pre-test:	4.68	1.7	m>0: t(70)=23.19, p<6.9e-34
Post-Test:	5.14	1.5	m>0: t(70)=28.17, p<3.19e-39
Raw Gain:	0.46	2.21	m>0: t(70)=1.75, p<0.04
NLG:	- 0.02	0.53	m#0: t(70)=-0.45, p<0.67, power = 0.02
Even-Grading Rubric			
Score	**Mean**	**sd**	**Comparison with baseline 0**
Pre-test:	3.79	1.82	m>0: t(70)=17.51, p<1.37e-26
Post-Test:	4.43	1.76	m>0: t(70)=21.18, p<1.83e-32

Raw Gain:	0.64	2.4	m>0: t(70)=2.24, p<0.02
NLG:	-0.02	0.44	m#0: t(70)=0.39, p<0.69, power = 0.05

As a consequence, while it is clear that some students gained from the experience, it is not clear that the gain is evenly spread across the entire student population. Moreover, as indicated by the normalized learning gain scores, while many of the students improved, the population as a whole did not achieve substantial learning gains despite the fact that, given their pre-test scores, they had room for significant improvement.

C. Performance Variance by Subgroups

To determine the impact of students' incoming skills and demographic characteristics, we assessed student performance by gender, undergraduate GPA, and LSAT. The class population was insufficiently diverse for us to assess the impact of students' ethnicity in a statistically meaningful way. Of the students in the course, forty-eight listed their ethnicity as "Caucasian/White," while thirteen declined to release the information, leaving only ten students of other ethnicities.

By contrast, the class breakdown was much more balanced in terms of gender, with thirty female students and forty-one male students. According to our analyses, the students' gender was not a significant factor in their performance either on the individual tests or in terms of learning gains. This was true for both the basic- and even-grading metrics.

To test the predictiveness of the students' undergraduate GPA, we performed a correlation analysis using Spearman's "rho"[80] statistic, which measures the correlation of linear valued

[80] Spearman's rho is a nonparametric test of correlation between two independent variables. A nonparametric test does not require a linear correlation. Here, the test for correlation is designed to test whether the correlation is 0, indicating no relationship between the values, or whether a significant correlation exists with rho being below 0 (indicating negative correlation) or above 0 (positive correlation). For example, in our study a rho value of .29 for the LSAT variable indicates that LSAT scores are positively correlated with higher post-test scores. Because this is a nonparametric test, however, this does not mean that the correlation between these variables is linear. *See* Dalgaard, *supra* n. 75; Myles Hollander & Douglas A. Wolfe, *Nonparametric Statistical Methods* 185–194 (John

variables. The relative correlation between the undergraduate GPA and pre-, post- and gain scores was low under both the basic and even-grading metrics. The strongest observed correlation was between the undergraduate GPA and students' post-test scores under the basic grading metric: rho=-0.1, r(69)=66157.23, p<0.37. This correlation, however, like the others was not significant. Therefore, undergraduate GPA was not a reliable predictor of student performance.

The LSAT score was a more useful predictor of student performance. Under the basic grading metric we observed predictive correlations between LSAT and the students' pre- and post-test scores (pre-test: rho=0.22, r(69)=46465.8, p<0.064; post-test: rho=0.29, r(69)=42460.26, p< 0.015). Intriguingly, no trend or significant correlation was observed between the LSAT score and students' raw gain or normalized learning gain (raw gain: rho=0.02, r(69)=58029.87, p<0.83; NLG: rho=0.017, r(69)=58630.17, p<0.89). This pattern was repeated with the even grading metric (pre-test: rho=0.25, r(69)=44763.46, p<0.04; post-test: rho=0.33, r(69)=39779.42, p<0.005; raw gain: rho=0.013, r(69)=58843.77, p<1; NLG: rho=0.004, r(69)=59388.05, p<1). These results indicate that, while the LSAT is predictive of students' basic competence, it does not signal the degree to which students will benefit from the traditional case-dialogue method. These results also may indicate that students are not taught but rather are simply ranked by the legal education process, or in the alternative, that our existing test metric is insufficiently sensitive.

D. Performance Variance by Question Type

In the tests we employed, three of the ten questions were single-case, determinate questions while the remaining questions were cross-case, indeterminate. We therefore applied our grading rubrics to the separate assessment of these question types with the goal of determining whether the students performed differently or improved overall as to each question type.

Wiley & Sons 1973).

1. Questions 1–3

Questions 1–3 were all single-case, determinate, single-answer questions. Thus, both grading metrics graded the questions the same with a maximum score of 3 for the set. Under these analyses the overall population scores were significantly greater than 0 on the pre-tests (m=1.42, sd=0.99, t(70)=12.04, p<5e-19) and on the post-tests (m=1.39, sd=1.02, t(70)=11.5, p<4.3e-18). The raw gain scores were not significantly different from the baseline of 0 under a two-sided t-test, nor were the normalized learning gains (raw gain: m=-0.03, sd=1.41, t(70)=-0.17, p<0.57, power = 0.035; NLG (m=-0.02, sd=0.17 t(70)=-0.83, p<0.4, power = 0.13).

Thus, we showed no major improvement in students' performance on single-case, determinate questions, but given their already high initial scores and the small sample size, our findings lack the statistical power to argue that no relevant learning took place.

2. Questions 4–10

Questions 4–10 were cross-case, indeterminate questions and included multi-choice questions, thus there was score variation between the two grading metrics. Under the basic grading rubric, the overall population scored significantly greater than 0 on the pre-test (m=3.25, sd=1.4, t(70)=19.58, p<2.05e-30) and on the post-test (m=3.74, sd=1.28, t(70)=24.56, p=1.99-36). In this case, the population showed significant raw gains (m=0.49, sd=1.68, t(70)=2.45, p=0.009). However, the normalized learning gain was not significantly different than the baseline under a two-sided t-test (m=0.04, sd=0.27, t(70)=1.16, p<0.25, power = 0.2). Again, the low power[81] of these tests prevents us from confirming a true no-learning-gain result.

Under the even-grading rubric, the overall pattern varied slightly. Again the pre- and post-tests were significantly greater than 0 (pre-test: m=2.37, sd=1.58, t(70)=12.64; p<4.79e-20; post-test: m=3.04, sd=1.52, t(70)=16.78, p<1.58e-26), as was the raw gain (m=0.67, sd=1.91, t(70)=2.93, p<0.003). In this case, however, we also found a trend (p<0.1) indicating that the normalized

[81] For a discussion of "power," see *supra* n. 76.

learning gain scores also rose significantly from the baseline: m=0.05, sd=0.27, t(70)=1.66, p<0.06. This increase was quite minor, on the order of making one fewer mistake on a multi-choice problem. Thus, while it is an undeniable gain, it is not substantial.

E. Performance Variance by Course Grades

As a final comparison we assessed whether the students' pre-, post-, and gain scores correlated with their final exam performance in the class. No significant correlation was found between the test scores or gain scores and the course grades. This suggests either that the skills reflected on the tests do not constitute a predictive part of the students' final exam performance or that the students' capacity to perform the skills measured in the tests changes over the final half of the semester.

IV. DISCUSSION

This study developed and utilized two test instruments to assess law student reading and reasoning skills. Through the comparison of two groups of students who were similar in terms of LSAT scores and who completed the tests in different order, the study determined that the test instruments are equivalent. In terms of both test performance and learning gains, there were no statistically significant differences between the groups. In addition, there were similar patterns of within-group differences, further indicating the equivalence of the test instruments.

In terms of performance for the study population as a whole, there were statistically significant raw learning gains. However, there were no statistically significant normalized learning gains. Overall, 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. This finding is consistent with prior studies that have found a lack of significant learning gains in terms of law student reading and reasoning skills.[82] These consistent findings may indicate that all of the test instruments are flawed, failing to measure actual gains in reading and reasoning skills. However, if the tests constitute accurate measures of these skills, the results across studies call for the development of teaching approaches other than the traditional case-dialogue method and for the continued assessment of learning gains.

The study found that gender and undergraduate GPA were not significant factors in predicting either performance on the tests or learning gains. LSAT scores were a significant factor in predicting performance on both the pre-tests and the post-tests, but not in predicting learning gains. Thus, LSAT scores appear to indicate the level of students' basic reading and reasoning skills. However, they do not appear to provide an indication of the degree to which students are likely to benefit from traditional law school instruction. In other words, the results of this study raise the question of whether LSAT scores fail to reliably predict law students' capacity for intellectual growth within the traditional legal education setting.

The study analyzed the results separately for each of the two question types—single-case, determinate and cross-case, indeterminate. Evensen's and Stratman's studies had indicated that these question types may test separate reading and reasoning skills.[83] For the single-case, determinate questions, the results indicated that there were no significant raw learning gains or normalized learning gains. For the cross-case indeterminate questions, there were significant raw learning gains under both the basic-grading rubric and the even-grading rubric. In addition, while there was not a significant normalized learning gain under the basic-grading rubric, there was a significant, albeit minor, normalized learning gain under the even-grading rubric. These divergent results as to learning gains provide some additional support for Evensen's and Stratman's finding that these two question types test different skills. However, given the relatively small number of single-case determinate questions, further study is necessary to validate this. In addition, the minor, but significant normalized learning gain for the cross-case, indeterminate question type under the even-grading rubric may indicate that

[82] Ashley, *supra* n. 10, at 350, 353; Evensen et al., *supra* n. 10, at 6, 27.

[83] Evensen et al., *supra* n. 10, at 7–8, 19.

traditional law instruction yields a minor positive effect on this most complex reading and reasoning skill.

The students' performance on the tests and their learning gains did not correlate with their performance on the final examination for the Legal Process course. This result may simply indicate that the students' legal reading and reasoning skills change significantly in the latter half of the semester. This result may also indicate that the level of reading and reasoning skills changes when students confront doctrinal areas other than personal jurisdiction. (The final examination required students to read new material and reason only in the area of applicable law in the federal courts in diversity of citizenship cases—i.e., the *Erie* doctrine.) This result could also be the product of the difference between multiple-choice tests and an essay examination, although both were designed to assess basic reading and reasoning skills. More specifically, this result could be a product of the difference in degree of subjective assessment and teacher error. In other words, it is possible this result indicates the instructor's incapacity to assess basic reading and reasoning skills in the grading of a final exam essay. Another possibility is that the final exam measures reading and reasoning skills that are different from the cross-case/hypothetical reasoning skill that the multiple-choice tests were designed to measure. In the alternative, this result could indicate that either the study's multiple-choice tests are not sensitive enough to accurately assess student reading and reasoning skills or the final exam is flawed in terms of measuring cross-case/hypothetical reasoning skill.

For most of the results discussed above the basic and even grading rubrics showed little, if any variation. However, we are encouraged by the results of the even-grading metric and the sensitivity of the analysis relative to the basic-grading rubric. As we discussed in the results section, it is indicated that students will err on the side of over-selecting responses and will thus be unfairly advantaged by a grading metric that does not penalize guessing. This being the case, we plan to apply the even-grading rubric in future studies.

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. Therefore, the study's findings are not 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 this regard, it should be noted that the instructor who taught the Legal Process course is an experienced, competent legal educator who regularly utilizes the traditional case-dialogue method of instruction. Also, the law school that participated in the study provides a traditional first-year curriculum and enrolls students who have qualifications that generally match those of students at schools ranked by *U.S. News & World Report* as being in the top 51–100 law schools in the country. Finally, the findings of this study are largely consistent with the findings of prior studies, especially in terms of a failure to find any significant learning gains in basic legal reading and reasoning skills.[84]

V. CONCLUSION

In summary, the findings of the studies in this area to date indicate that, while legal education does not diminish law students' skills of reading and reasoning, neither does it significantly enhance these skills. The results of these studies provide a foundation for future studies that introduce and measure the effect of new educational interventions. As discussed in this Article, Ashley's studies have begun to pursue this line of inquiry.[85] Studies of legal reading strategies have also pursued this line of inquiry, finding that law students' reading skills can be enhanced through educational interventions, such as checklists (e.g., providing reading guidelines for case analysis that set out detailed tips for putting the case in context, reading the case for an overview, and rereading the case analytically) and role assignments (e.g., placing students in a particular attorney role—advocate, advisor, policy analyst—for a reading task).[86] Thus, there is hope that educa-

[84] *See supra* nn. 20–64 and the accompanying text.

[85] *See* Ashley, *supra* n. 10.

[86] *See* Mary A. Lundeberg, *Metacognitive Aspects of Reading Comprehension: Studying Understanding in Legal Case Analysis*, 22 Reading Res. Q. 407, 429 (1987) (finding that the provision of reading guidelines significantly improved legal reading comprehension); James F. Stratman, *When Law Students Read Cases: Exploring Relations Between Professional Legal Reasoning Roles and Problem Detection*, 34 Discourse Processes 57, 57, 60–62, 64–68 (2002) (finding that law students comprehended text better when they as-

tional interventions could enhance learning gains in the basic skills of legal reading and reasoning. We plan to pursue this line of inquiry by introducing a writing exercise that provides students with formative feedback at the mid-point of the personal jurisdiction unit. (Our next paper will report the results achieved by this educational intervention.) We will then introduce additional educational interventions one-by-one (e.g., the use of think aloud methodology to generate class discussion; the use of social media-type software for small group reading and annotation of legal texts) and measure their effects on learning outcomes.

sumed the role of an actual attorney rather than simply reading to prepare for class discussion); *see also* Leah M. Christensen, *Legal Reading and Success in Law School: An Empirical Study*, 30 Seattle U. L. Rev. 603 (2007).