ASSESSING DIFFERENTLY AND USING EMPIRICAL STUDIES TO SEE IF IT MAKES A DIFFERENCE: CAN LAW SCHOOLS DO IT BETTER?

Andrea A. Curcio*

I. INTRODUCTION

From the LSAT through law school classes and ending with the bar examination, the methods of assessing potential lawyers involve testing a very narrow range of skills and assessing those skills within an equally narrow range of test methodologies. These assessment methods have repeatedly been critiqued as an inadequate and inaccurate way to develop and assess the skills and values that new lawyers need to practice law competently. It is now time to move beyond the critiques and toward a scholarly exploration of whether law professors can develop valid, reliable, and cost-effective assessments that better test for a wider range of the skills competent lawyers need, and whether they

---

* Professor of Law, Georgia State University College of Law. Thank you to Clark Cunningham, Dr. Harry Dangel, Dr. Gregory Jones, Eileen Kaufman, Steven Kaminshine, Jonathan Todres, and Hazel Weiser for their helpful comments on earlier drafts of this essay. A special thanks to the incredible Georgia State University College of Law Research Librarians for their assistance with the social science research.


3. For a discussion of the range of skills lawyers need, see, e.g., Schultz & Zeck,
can design assessments that integrate, rather than segregate, legal analysis, practical skills, and professional values. This Article lays the groundwork for exploration of those issues by proposing concrete alternative assessment ideas, briefly explaining how to develop empirical studies to measure the validity and reliability of those and other assessments and, finally, suggesting various potential empirical assessment studies.

One obvious reason to explore alternative assessment methods is to ensure that law schools produce the most effective lawyers possible. Less obvious—but equally compelling—reasons to look at alternative assessments are that existing assessments have significant consequences in terms of the profession’s diversity, students’ job opportunities, and students’ mental health.

There is no question that current law school assessment methods and their bar exam counterparts serve as a barrier to entry into the profession for many people of color. The LSAT is heavily relied upon as a law school admissions tool because of its correlation to law students’ grades. Yet, the emphasis on LSAT scores in law school admissions decisions “substantially reduces the presence of African-Americans and Latino students in law school and the profession, as well as diminishing the prospects of inclusion for those from most non-elite families.” Students of color who make it through the LSAT hurdle have, on average, a higher first-year dropout rate, lower first-year and final law school GPAs, and lower bar passage rates than their white classmates.

supra note 1, at 26-27.

4. The Carnegie Report emphasized the need to eliminate the dichotomy between doctrinal analysis and skills training, as well as the need to integrate professional identity development into the law school curriculum. See SULLIVAN ET AL., supra note 2, at 31, 88-89.


In addition to their negative impact upon diversifying the profession, law school grades and class rankings also have an effect upon law graduates' initial job opportunities. Finally, law school grades also impact students' sense of confidence and self-worth, and often cause students to disengage. Disturbingly, law school assessments do all this without ever having been proven to measure validly and reliably who will actually be an effective lawyer.

In an extensive empirical study in which 2,000 lawyers were surveyed, Professors Schultz and Zedeck identified twenty-six factors important to effective lawyering. These include factors such as: "analysis and reasoning; creativity and innovation; problem solving; practical judgment; researching the law; fact finding, questioning and interviewing; influencing and advocating; writing; speaking; listening; negotiation skills; the ability to see the world through the eyes of others;"

additional empirical analysis, which demonstrates that students of color have a higher first-year dropout rate, lower first-year and final law school GPAs, and lower bar passage rates than their white counterparts; see also LINDA F. WIGHTMAN, LSAC NATIONAL LONGITUDINAL BAR PASSAGE STUDY 27 (1988) (finding significant difference in bar pass rates between white and minority first-time bar exam takers); William C. Kidder, The Bar Exam and a Dream Deferred: A Critical Analysis of the MBE, Social Closure, Racial and Ethnic Stratification, 29 LAW & SOC. INQUIRY 547 (2004) (discussing the negative impact of raising bar exam cut scores on minority bar applicants).


9. For an excellent review of the literature on the impact of law school grades on students' mental health, see Grant H. Morris, Preparing Law Students for Disappointing Exam Results: Lessons from Casey at the Bat, 45 SAN DIEGO L. REV. 441 (2008).

10. See id. at 452-53 (discussing how low grades cause students to disengage from the learning process).

11. See, e.g., STUCKEY ET AL., supra note 2, at 236-39 (critiquing the predominant existing methods of assessing law students); Greg Sergienko, New Modes of Assessment, 38 SAN DIEGO L. REV. 463, 468-75 (2001) (discussing reliability and validity problems with existing law school assessment methods). In addition to problems with validity and reliability of current assessment methods, the assessments themselves are so narrow in scope that many factors that predict effective lawyering are not evaluated. See SCHULTZ & ZEDECK, supra note 1, at 24-27 (describing process of determining factors important to effective lawyering, many of which are not assessed by the LSAT, law school exams, or the bar exam).

12. See SCHULTZ & ZEDECK, supra note 1, at 26-27.
diligence;,, and self-development.”\textsuperscript{13} These are similar to the practical skills and ethical components of a competent lawyer that were identified twenty-five years ago in the MacCrate Report.\textsuperscript{14} Yet, virtually all assessment methods from the LSAT through the bar exam measure only a very small percentage of the skills and qualities necessary to be an effective and competent attorney. In fact, the lack of a coherent structure and plan to teach and assess many of these practical skills and professionalism qualities is a central critique of the recent Carnegie Report on the Future of Legal Education\textsuperscript{15} and the Best Practices book.\textsuperscript{16}

Not everyone agrees that assessments should change. Some argue that multiple choice, short essay, and longer essay question assessment methods may not be perfect, but they are a good measure of the baseline skills all lawyers need and they are a valid, reliable, and cost-effective way to measure those baseline skills.\textsuperscript{17} Even if one accepts that argument, however, it does not justify inertia. Until law professors explore alternatives, they will not know what is possible. For example, a comprehensive empirical study by Professors Schultz and Zedeck indicates that it may be possible to modify or supplement the LSAT to include “lawyer effectiveness” measure predictors that greatly reduce or eliminate the LSAT’s discriminatory impact, while retaining its validity as a law school admissions tool.\textsuperscript{18} That same kind of exploration of alternative assessment measures should be occurring, on a much smaller scale, in law school classrooms across the country. Without

\textsuperscript{13} Id. Schultz and Zedeck also identified the following skills: strategic planning; organizing and managing one’s own work; organizing and managing others (staff/colleagues); networking and business development; providing advice and counsel, and building relationships with clients; developing relationships within the legal profession; evaluation, development, and mentoring; passion and engagement; integrity and honesty; stress management; and community involvement and service. Id. These are also some of the factors identified by others as necessary to good lawyering. See, e.g., Stuckey et al., supra note 2, at 51, 78.

\textsuperscript{14} AM. BAR ASS’N, LEGAL EDUCATION AND PROFESSIONAL DEVELOPMENT—AN EDUCATIONAL CONTINUUM, REPORT ON THE TASK FORCE ON LAW SCHOOLS AND THE PROFESSION: NARROWING THE GAP 138-41 (1992) [hereinafter MACCRATE REPORT].

\textsuperscript{15} Sullivan et al., supra note 2.

\textsuperscript{16} Stuckey et al., supra note 2.

\textsuperscript{17} See, e.g., Lorenzo A. Trujillo, The Relationship Between Law School and the Bar Exam: A Look at Assessment and Student Success, 78 U. Colo. L. Rev. 69, 85-87 (2007) (recounting arguments supporting the contention that the bar exam is a good indicator of who will be a competent new lawyer); Suzanne Darrow-Kleinhaus, A Response to the Society of American Law Teachers Statement on the Bar Exam, 54 J. LEGAL EDUC. 442, 443 (2004) (arguing that the bar exam “appropriately serves its purpose”).

\textsuperscript{18} Schultz & Zedeck, supra note 1, at 55-58.
experimentation and study, law professors will never know if they can develop valid, reliable, and easy-to-administer assessments that measure a wider range of lawyering skills or even if there are ways to improve the reliability and validity of existing assessment methods.\(^{19}\)

This Article explores alternative law school assessment methods and suggests ways to engage in empirical study of alternative and existing assessments. In Part II, the Article provides some concrete suggestions for both incremental and larger changes to doctrinal class assessment practices with an eye toward developing assessments that begin to desegregate legal analysis, practical skills, and professional identity.\(^{20}\) What is not explored explicitly, but what is implicit throughout, is that, because professors should not assess what they do not teach, examining and changing assessment methods necessarily involves reflection about teaching methodology and coverage.\(^{21}\)

The Article then moves from a discussion of changing teaching and assessment practices to a discussion of studying the impact of those changes. Because empirical studies can be a persuasive tool in laying the groundwork necessary to develop institutional support for a change in assessment practices, Part III seeks to de-mystify the empirical study process. Written for those without a social science background, it briefly discusses basic issues in study design, methodology, implementation, and interpretation, providing an overview of how to develop and design an empirical research study. Part IV suggests some potential empirical research assessment studies that can be performed on both alternative and existing law school assessment methods. Finally, the Article concludes by arguing that if law professors give their teaching and assessment work the same scholarly scrutiny given to other research interests, they may discover ways to help students become more effective lawyers, lessen student disengagement and create the potential for a more diverse bench and bar.


\(^{20}\) The Carnegie Report calls for better integrating legal analysis, practical skills and professional identity. See SULLIVAN ET AL., supra note 2, at 194-200.

\(^{21}\) See infra text accompanying notes 125-29 (discussing assessment of construct and content validity).
II. ASSESSING DIFFERENTLY: INCORPORATING THE CARNEGIE MODEL INTO DOCTRINAL COURSE ASSESSMENTS

This section briefly discusses a few ways that doctrinal professors can incorporate a broader range of assessments into their courses, and ways that those assessments can better integrate the cognitive, the practical, and the professional identity dimensions of lawyering. It does so recognizing that in order to assess different skills, one must first teach those skills, which may reduce the amount of substantive law coverage in a given class. It also acknowledges that the development of alternative assessment methods will increase the time law professors must devote to teaching and assessing, and thus may impact the time available to engage in legal scholarship, which is the "coin of the realm" in terms of professorial status and career advancement. Although these are significant concerns, they are not insurmountable barriers. Thus, with recognition that developing alternative assessments involves

22. The Carnegie Report describes the need for law schools to integrate better the three professional apprenticeships. See SULLIVAN ET AL., supra note 2, at 191-200. The first apprenticeship is the cognitive, which essentially is the teaching and learning of legal reasoning and analysis. Id. at 28. The second apprenticeship is the practical, which is the development of skills traditionally taught in clinics and simulation courses, such as fact development, interviewing, counseling, oral and written advocacy, and legal drafting. Id. The final apprenticeship is that of professional identity and purpose, which encompasses the development of an understanding of lawyers' ethical duties, responsibilities, and the lawyers' role as an advocate for justice. Id.

23. See infra text accompanying notes 128-29 (noting that an assessment does not have content validity if the skill sought to be assessed has not been taught). Ideally, skills will be taught across the curriculum rather than in a specific course. For a discussion of how an Australian law school integrated skills teaching throughout its curriculum, see Bobette Wolksi, Why, How and What to Practice: Integrating Skills Teaching and Learning in the Undergraduate Law Curriculum, 52 J. LEGAL EDUC. 287 (2002).


25. Studying assessments can result in publishable scholarship, although it may not be credited in the same way by colleagues as more esoteric law review articles. If there are such studies, however, and they indicate that alternative assessment methods improve student learning, produce more effective lawyers, or eliminate the discriminatory impact of existing assessment methods, the studies may help persuade law schools that they owe students a fiduciary responsibility to value teaching and assessing, and its resulting scholarship, to the same extent as more abstract legal scholarship, and the data may even create market demand for changing current practices.
trade-offs, the following subsection provides a starting point for those seeking specific alternative assessment ideas.

A. Factual Development Assessments

Factual development is a basic, yet critical, lawyering skill. Some factual development skills can be incorporated into virtually all doctrinal courses with only minor adjustments to existing teaching and assessment methodologies. In a wide variety of doctrinal courses, students could be asked to develop questions that tie fact gathering into legal rules application. For example, in a criminal procedure class, students could identify the questions they would want to ask the officer prior to a motion to suppress hearing. In a torts class, students could be asked to develop the questions they would need answered in order to determine if a person was acting in the course and scope of her employment for respondeat superior liability.

Virtually every professor could, with relative ease, devise problems in which more facts were necessary to analyze fully the legal doctrine. This factual development skill could be practiced in class and assessed on an exam. For example, in class, and then on an exam, students could be asked questions such as: “What facts are necessary in order to analyze fully this problem, and how would you go about finding those facts?” In essence, incorporating basic fact gathering into teaching and assessment simply involves analyzing existing cases and problems through a slightly different lens. 

26. For example, on an evidence exam, the author asked the following question, which derived from a problem in which a plaintiff was injured while working on a machine, and the defense was that the plaintiff was contributorily negligent:

As required by law, the police department investigated an accident, and the investigating officer wrote a report. In the “witnesses” section of the report, it states: “Jerry Bell [a foreman] reported that shortly before plaintiff was injured, the machine’s surface was covered with debris, as well as slime and water.” If plaintiff objected to this section of the report on hearsay grounds, what information does the judge need before ruling on this issue? Does your answer change if it is the plaintiff who wants to admit the report, and the defendant objects on hearsay grounds?

To answer this question, students needed to know the hearsay rules and, then, had to think about how the facts could apply to those rules. For example, when the plaintiff objected, students should have been thinking first about whether this fits under the present sense impression or excited utterance exception, and what additional facts the judge would need to determine if it was a present sense impression—e.g., did the foreman have firsthand knowledge of the machine’s surface or was that told to him by someone else? How soon after the fall was the statement made?; What was the foreman’s demeanor at the time he made the
Actual factual investigation and development in the context of legal doctrine also can be fairly easily assessed. For example, civil procedure professors could draft a take-home exercise in which students must conduct independent research to determine a corporate defendant's state of incorporation and principal place of business in order to perform a diversity analysis. In a property or real estate transactions class, the professor could devise a problem, the answer to which depends on whether and when a deed has been recorded. The students may be required to conduct online research or go to the courthouse to determine whether the deed was properly recorded. In a corporations class, the professor could require students to pull documents from the secretary of state's website or office in order to answer a particular question. With a little imagination, or consultation with practicing attorneys, doctrinal professors can identify basic factual research issues arising in practice which could be performed by the students either through online or courthouse research.

B. Assessments via Video Reviews of Doctrinal Course Simulation Exercises

One way to assess listening, problem solving, ethical issue identification, and client counseling skills as they relate to doctrinal concepts is to develop in-class videotaped simulations which require out-of-class analysis. For example, in an employment discrimination course, a professor could tell students that a client with a potential racial discrimination claim was coming in for an interview. Or, students could be given a fact scenario involving a sexual harassment claim and told to be prepared to counsel the client. Two teams of students could perform an in-class videotaped simulated client/lawyer meeting. The professor could post the video on a course website and require students to write a short paper analyzing how well the student "lawyers" identified the client's objectives, listened to the client's concerns, saw the world through the eyes of their client, identified any embedded ethical issues, and suggested potential courses of action available to the client using the applicable doctrine as it applied to the facts of the particular case.27

________________________
27. This idea is taken from Clark D. Cunningham, "How Can We Give Up Our Child?"
C. Document Drafting Assessments

Document drafting exercises provide a good opportunity to assess students on doctrine, skills, and professionalism. In most substantive classes, students' understanding of the rules of law and their application to the facts of a given case could be examined through a complaint-drafting exercise that requires students to allege the predicate facts for each element of a claim, as well as set out the applicable claim elements.28

In a transactional class, a professor could assess students' substantive knowledge through a contract clause drafting or editing problem.29 Wills and trusts courses present natural opportunities for drafting assessments, as do many other transactional courses. In all these courses, ethical and professionalism issues could be embedded into the drafting problem.30

The main problem with drafting assessments is the time it takes to grade students' work, especially in large-section courses. One solution to this dilemma may be to assign a narrow range of grades (e.g., ten points for an excellent document, eight points for a good one, and six points for one that needs significant improvement). Then, assuming a well-developed rubric, and training in how to apply the rubric, research assistants could potentially handle the grading,31 thus alleviating some of the time constraint issues that might otherwise deprive students of these drafting assessment opportunities.

---

28. To make the exercise more challenging, a "client" could be interviewed by the professor or by student teams and the complaint would have to be drafted using information gathered in that interview. A rubric for this kind of exercise could allocate a few points for organization, but the majority of allocated points would be for demonstrated applied knowledge. For more information about developing this kind of exercise and accompanying rubric, see Andrea A. Curcio, Moving in the Direction of Best Practices and the Carnegie Report: Reflections on Using Multiple Assessments in a Large-Section Doctrinal Course, 19 Widener L. J., (forthcoming 2009).

29. For a discussion on how to do this, see Deborah A. Schmedemann, Finding a Happy Medium: Teaching Contract Creation in the First Year, 5 J. ASS'N LEGAL WRITING DIRECTORS 177 (2008) (discussing integrating contract clause drafting into a legal writing course).

30. For some examples of how to do this, see discussion infra, Part II.E.

31. Using graduate assistants to grade exams and papers is a common practice in most undergraduate research institutions.
D. Oral Performance Assessments

Oral communication skills are indisputably a key component of effective lawyering. Yet, in most doctrinal courses, students seldom are assessed on their ability to integrate issue spotting and legal analysis into an oral communication format. Professors could utilize simulated client interviews or counseling exercises to assess students’ ability to analyze and orally communicate their analyses. The most sophisticated way to do this is to use the “standardized client” method, which was empirically tested and validated in Scotland and is currently being used in the Daniel Webster Scholar Program at Franklin Pierce Law School. The standardized client model uses trained actors to play the role of clients and then assess the students’ performance. The “client” could assess the interviewing skills based upon the well-researched and tested rubric developed by Barton and her colleagues, and the professor could assess the students’ grasp of the doctrine. The time and expense of this type of assessment exercise may make it more suitable to upper-level smaller-section doctrinal courses.

Motions arguments—involving doctrinal issues arising in the course—provide an oral-assessment exercise suitable even for larger-section courses. These could potentially be done by training research assistants or alumni to be the “judges.” The arguments could be videotaped and later reviewed using a rubric similar to one professors use when grading an essay exam question. The rubric should be altered slightly, however, to provide points for clarity of presentation rather than clarity of writing. Students could even be asked to review their own

32. For an example of how to design a law student oral assessment exercise, see John M. Burman, *Oral Examinations as a Method of Examining Law Students*, 51 J. LEGAL EDUC. 130 (2001).
35. Barton et al, supra note 33, at 58.
36. To be a valid assessment, students should be instructed on how to make a motions argument. See infra text accompanying notes 128-29 (discussing content validity). Also, to make the assessment a useful learning exercise, students should be given an opportunity to practice this type of argument before a graded assessment. See STUCKEY ET AL., supra note 2, at 255-59 (suggesting that formative assessments throughout the semester improve student learning).
video and complete a self-critique.  

E. Assessing Ethical Decision-Making and Professional Responsibility

In general, outside of professional responsibility classes, students seldom are assessed on their ability to identify and address ethical issues. This can change without radically altering a professor's current teaching style. In virtually any doctrinal course, a professor can develop problems that raise ethical or professionalism issues. Likewise, even in the most traditional essay exams, students could be asked to identify and discuss any ethical issues they see. If professors are engaging in alternative assessment exercises, ethical and professionalism issues can be built into the problems. For example, in a complaint or answer drafting exercise, students can be confronted with a client conflict issue or a potential Rule 11 violation problem. In an oral argument assessment, students could be told that their client does not want them to disclose particular information that the student knows would persuade the judge, thus presenting the student with a potential professionalism issue.

Ultimately, assessments should embed ethical dilemmas and professionalism issues, and students should be assessed on their ability to recognize and resolve these issues in virtually every law school class. Of course, the ability to do this depends upon students' exposure to the applicable ethical rules. If the students have not yet taken a professional responsibility course, the applicable rules should be taught and discussed in class before students are assessed on their ability to identify and analyze ethical issues.

F. Assessing Students' Ability to See the World Through the Eyes of Others

Another effectiveness factor identified by Professors Schultz and Zedeck is the ability to see the world through the eyes of others. Yet,

38. Rule 11 prohibits attorneys from alleging a claim or asserting a defense unless they have a good faith belief that evidence exists to support that claim or defense. FED. R. CIV. P. 11(b).
39. See infra text accompanying notes 128-29 (discussing content validity).
40. SCHULTZ & ZEDECK, supra note 1, at 26.
outside of the clinical education context, this essential component of lawyering is seldom taught, let alone assessed. That can change. In various exercises and papers, students could be evaluated based on their ability to identify, practice, and reflect upon the types of attitudinal traits that help facilitate effective interactions with people with diverse backgrounds. These traits could include, among other things, such qualities as understanding of the reasons why a person might say/act/react in a particular way, openness to a different way of viewing the world, personal self-awareness, and cultural self-awareness. If faculty members adopt the assessment suggestions set forth in Part II.B, students, in reviewing a videotaped simulated client meeting, could be asked to consider such questions as: “What message was the lawyer’s body language sending to the client?”; “Did the lawyer conduct the meeting so as to encourage the client to tell her story and explain her goals?”; “How does your world-view and upbringing influence your view of the client’s problem?”

G. Teamwork Assessments

Many of the assessments described above could be completed either individually or in teams. Collaboration enhances learning, and working well with others is a critical lawyering skill. Teamwork itself presents an opportunity for students to explore and experience professional identity and professionalism issues. Teamwork could be assessed both by other team members and through self-reflection by asking questions relevant to the students’ participation in the team effort.

41. STUCKEY ET AL., supra note 2, at 119-21 (discussing studies showing that collaboration enhances learning).

42. See, e.g., Margaret M. Barry, A Question of Mission: Catholic Law School’s Domestic Violence Clinic, 38 HOW. L.J. 135, 146 (1994) (discussing the importance of teaching group collaboration, cooperation with other professionals and consensus building); Steven L. Schwarcz, Explaining the Value of Transactional Lawyering, 12 STAN. J.L. BUS. & FIN. 486, 508 (2007) (noting the value of teaching students about the importance of cooperation with opposing counsel when doing transactional work); David B. Wilkins, A Systematic Response to Systemic Disadvantage: A Response to Sander, 57 STAN. L. REV. 1915, 1925 (2005) (suggesting that teamwork, creativity, and good judgment are critical components of good problem solving and successful lawyering).

43. For an excellent discussion of the value of how to design and implement a web-based template for self and peer assessment of teamwork contributions see Mark Freeman and Jo McKenzie, SPARK, A Confidential Web-Based Template for Self and Peer Assessment of Student Teamwork: Benefits of Evaluating Across Different Subjects, 33 BRITISH J. OF EDUC. TECH., 555 (2002). In evaluating teamwork contributions, professors can develop a template
H. Assessing the Alternatives

The above suggestions are not a comprehensive list but, rather, a starting point for thinking about assessments in light of the Carnegie Report's call to desegregate how and what we teach so that students' legal education allows them to understand how legal analysis, practical skills, and professional identity are intertwined.\textsuperscript{44} A question that arises with alternative assessments such as these (as well as with traditional assessments such as multiple choice, short answer, and essay questions), is whether they can be shown to be valid, reliable, and cost-effective. Empirical research presents one way to demonstrate these traits.

Empirical research can be used to gather information about the impact of assessments on student learning, to determine the validity and reliability of particular assessment methods, and to improve teaching. Information gleaned from empirical studies may cause law professors to re-design assessments, and this, in turn, may lead to re-thinking approaches to teaching the skills being assessed.\textsuperscript{45} Empirical research also has significant persuasive power and can potentially be used to persuade faculty, and even bar examiners, to re-examine the methodologies they have been using.\textsuperscript{46} Thus, professors should think not only about how they can design and implement alternative assessments, but also about the nature of the student's contribution, the timeliness of the contribution and the value of the contribution. For sample questions on these topics, as well as an excellent guide to designing a valid and reliable teamwork assessment model, see Timothy J. Ellis and William Hafner, Peer Evaluations of Collaborative Learning Experiences Conveyed Through an Asynchronous Learning Network, in \textit{Proceedings of the 38th Annual Hawaii International Conference on System Sciences} (2005) \textit{available at} csdl.computer.org/comp/proceedings/hicss/2005/2268/01/22680005b.pdf. Some questions that could be asked on a teamwork student and peer assessment template include: to what extent did the student keep up with assigned deadlines; to what extent was the student's work thorough and reliable; to what extent did the student contribute to the group's discussion and planning; what percentage of the completed project can be attributed to the student's individual efforts?\textsuperscript{44\textsuperscript{4}}

\textsuperscript{44.} See Sullivan et al., supra note 2, at 194.

\textsuperscript{45.} See, e.g., Barton et al., supra note 33, at 52 (noting that empirical assessment research led faculty at Glasgow Graduate School of Law to "re-think its whole approach to teaching and learning interviewing and client communication skills").

\textsuperscript{46.} Mary Sue MacNealy, Strategies for Empirical Research in Writing 36 (1999) (discussing the persuasive power of empirical work). For an example of a persuasive empirical study, see Schultz & Zedeck, supra note 1 (presenting powerful preliminary empirical evidence that the LSAT can be re-drafted to encompass strong predictors of lawyering effectiveness that do not have the same disparate impact on minority test-takers as does the current LSAT). If law professors demonstrate the viability of assessing a wider range of skills, it is possible that bar examiners may follow suit.

\textsuperscript{44.} See Sullivan et al., supra note 2, at 194.
assessments, but also about empirical examination of those alternatives to determine whether they are effective learning and assessment measures. Likewise, for those who believe that the current law school assessment methods work well and there is no need for change, empirical examination of those assessments might validate or repudiate that position. The following section provides a basic overview of the issues professors should think about when considering empirical exploration of assessment methods.

III. DEVELOPING AN EMPIRICAL ASSESSMENT-METHODS STUDY

This section provides an overview of the steps involved in designing and implementing an empirical research assessment study. It discusses how to identify the research question, how to design the study, the need for human subject research approval, implementation and data analysis issues, and how to report the study’s results. This section seeks to de-mystify the empirical study process for law professors interested in assessment research who have little or no statistical methodology experience or training. It assumes that law professors without a social science or statistical background will collaborate with statisticians or social scientists. Thus, it simply seeks to familiarize law professors unschooled in empirical research with basic issues in study design, methodology, implementation, and interpretation.

A. Identifying the Research Question

The initial step in designing an empirical study is more challenging than it may first appear: a researcher must identify the question to be answered. In any assessment study, developing the research question requires identification of exactly what the professor seeks to assess. It is at this point that identifying learning outcomes may help clarify a research question, especially if one wants to examine the impact of
different teaching or assessment methods on student learning. Learning outcomes involve demonstrable achievement or progress in terms of students’ knowledge, skills, or attitudes/values. The outcomes can be very broad (e.g., students will develop strong communication skills) or very narrow (e.g., when conducting an interview, students will know how to ask open-ended questions).

There are numerous sources to help researchers identify potential law student learning outcomes. Educating Lawyers: Preparation for the Practice, Best Practices for Legal Education, and the MacCrate Report provide a good starting point. Additional resources include the work of law professors and scholars from other countries who have developed comprehensive learning outcomes for their own students, and surveys of practitioners and judges seeking information about the overarching concepts that it encompasses.”

Ellington and Earl, Facilitating Student Learning: A Practical Guide for Tertiary-Level Teachers 28 (1999). Next, there are Objectives which are “sets of more detailed statements that specify the means by which the various aims of the course, course unit, package, etc. will be met, i.e [sic] relate to the activities that it involves and the content that it covers. In other words, objectives relate to the course, course unit, or package.” Id. Finally, there are Learning Outcomes, which Ellington and Earl describe as follows:

[Sets of even more detailed statements that specify the various things that students will be able to do after successful completion of the learning process that the course, course unit, package, etc. covers, i.e [sic] relate to the new skills and assessable changes in behaviour that it is designed to bring about. In other words, learning outcomes relate to the minimum and demonstrable achievement of the students.]

Id. For further discussion of the principles discussed above, see Declan Kennedy et al., Writing and Using Learning Outcomes, A Practical Guide 5-6 (2006), available at http://www.bologna.msmt.cz/files/learningoutcomes.pdf. For a comprehensive list of potential teaching goals and learning outcomes on the undergraduate level, many of which are equally applicable to law teaching, see Thomas A. Angelo & K. Patricia Cross, Classroom Assessment Techniques, A Handbook for College Teachers 20-22 (2d ed. 1993). For examples of how those goals can be translated into a classroom assessment project, see id. at 42-44.

49. For a list of the various definitions of learning outcomes, see Kennedy et al., supra note 48, at 4. For an in depth discussion on learning outcome measures for law students, see Stuckey et al., supra note 2, at 42-91; see also Gregory S. Munro, Outcomes Assessment for Law Schools (2000).

50. Sullivan et al., supra note 2.
51. Stuckey et al., supra note 2.
53. See, e.g., Stuckey et al., supra note 2, at 53-54 (setting forth learning outcomes of the Law Society of England and Wales); Sharon Christensen & Sally Kift, Graduate Attributes and Legal Skills: Integration or Disintegration? 11 Legal Educ. Rev. 207 (2000) (discussing the development of integration of skills across the law school curriculum and discussing how those skills relate to particular learning outcomes Australian law graduates should possess).
skills and qualities new lawyers should possess.\textsuperscript{54}

Learning outcomes for research purposes should contain the following three characteristics: "1. The specified action by the learners must be observable; 2. The specified action by the learners must be measurable; and 3. The specified action must be done by the learners."\textsuperscript{55} Learning outcomes focus on what the student can and cannot do. Thus, clearly delineating the learning outcomes sought to be produced and measured can help identify the research question.

Whether using learning outcomes or some other criteria,\textsuperscript{56} it is important to define a research question as narrowly and precisely as possible.\textsuperscript{57} For example, assume a professor wants to measure whether a student’s ability to listen actively improves throughout the semester following a series of exercises or other assignments designed to improve these skills. Rather than examine an improvement in listening skills per se, it might be better to design a study that focuses on a few traits of active listening, such as whether the student allows his/her partners to finish their thoughts before responding, or whether the student clarifies his/her understanding of the point made by re-stating or summarizing it.\textsuperscript{58} The more narrowly defined the question, the easier it will be to develop specific methods for empirical analysis.\textsuperscript{59}

\textsuperscript{54} For a list of those skills, see, e.g., SCHULTZ & ZEDECK, \textit{supra} note 1, at 26-27; John O. Sonsteng et al., \textit{A Legal Education Renaissance: A Practical Approach for the Twenty-First Century}, 34 WM. MITCHELL L. REV. 303, 371-77 (2007) (discussing results of surveys of practicing attorneys regarding lawyering skills that they thought were important).


\textsuperscript{56} For example, in developing the standardized client model, Barton and her colleagues looked at refining the interview and assessment exercise to focus on those components of client interviewing that could be evaluated accurately by non-lawyers. Barton et al., \textit{supra} note 33, at 4.

\textsuperscript{57} Richard K. Neumann, Jr. & Stefan H. Krieger, \textit{Empirical Inquiry Twenty-Five Years After the Lawyering Process}, 10 CLINICAL L. REV. 349, 356 (2003); see also Barton et al., \textit{supra} note 33 (discussing how they repeatedly narrowed their interview questions and assessment instruments).


\textsuperscript{59} MACNEALY, \textit{supra} note 46, at 77-79. For a good discussion of why an empirical research question must usually be narrow in order to be useful, see Neumann & Krieger,
B. Developing and Designing an Empirical Study

Law professors without a social science or statistical training background likely will need a social science collaborator to help design, implement, and analyze an empirical study. There are numerous potential avenues to identify collaborators. University centers for teaching and learning, education departments, and business schools may be fruitful places to look for collaborators. Another good place to identify collaborators is through a literature search—identifying others who have done similar work in different disciplines and who may be willing to collaborate on a similar study with law students.

Conducting a literature search to identify other studies that address similar assessment research questions not only helps identify potential collaborators, it is also a critical step in developing an empirical study. There is a small, but growing, body of empirical research on law school teaching methodologies and assessments. Additionally, other disciplines have been investigating teaching and assessment methodology for a long time, and those studies can provide useful

supra note 57, at 372-74.

60. In addition to these resources, the Society of American Law Teachers [SALT] anticipates that by the spring of 2010, it will have compiled a list of potential assessment research social science collaborators that will be posted on its website, www.saltlaw.org.


62. Two disciplines that have done substantial work on assessments are business and medical education. A good resource for assessment research in business education is ASSESSMENT OF STUDENT LEARNING IN BUSINESS SCHOOLS: BEST PRACTICES EACH STEP OF THE WAY (Kathryn Martell & Thomas Calderon eds., 2005). A good starting point for looking at assessment research in medical schools is to review the discussion of assessment research and citations to assessment studies contained in Ronald M. Epstein, Assessment in
Performing a literature search before beginning also will help in terms of study design and later situating the data analysis within the existing literature. Additionally, it might generate ideas the researcher had not considered. Social science research is built upon others' work in a very visible manner. Thus, there are no pre-emption worries and it is perfectly valid to repeat a study done by someone else to see if a second study produces similar results.

After identifying a research question and collaborator, the next step is to choose a study method. Three essential characteristics of good empirical research are that the study is planned in advance of the data collection; the data is collected systematically, and the data collection produces evidence that can be examined by others. Planning the study requires determining in advance which method or combination of

---

Medical Education, 356 NEW ENG. J. MED. 387 (2007). For a sample medical student assessment study, see, e.g., Patricia J. Foster, Verbal Participation and Outcomes in Medical Education: A Study of Third-Year Clinical Discussion Groups, in STUDIES OF COLLEGE TEACHING 117 (Carolyn L. Ellner & Carol P. Barnes eds., 1983) (studying the relationship between medical students' verbal participation and their performance on standardized critical thinking skills and medical knowledge evaluations).

63. Some potentially useful studies in other disciplines that may provide models for similar studies of law students include: Daryl G. Smith, Instructions and Outcomes in an Undergraduate Setting, in STUDIES OF COLLEGE TEACHING 83 (Carolyn L. Ellner & Carol P. Barnes eds., 1983) (discussing the relationship between the processes of instruction and students' development of critical thinking skills); Linda Bol & Douglas J. Hacker, A Comparison of the Effects of Practice Tests and Traditional Review on Performance and Calibration, 69 J. EXPERIMENTAL EDUC. 133 (2001) (studying the impact of practice tests on graduate students in an introductory research-methods-in-education course and including an extensive literature review of similar studies); Jeanne Cameron et al., Assessment as Critical Praxis: A Community College Experience, 30 TEACHING SOC. 414 (2002) (discussing study of the relationship between summative and formative assessments in introductory sociology courses, using quantitative (student survey) and qualitative (faculty observations) empirical study methods); Kirsten Crossgrove & Kristen L. Carron, Using Clickers in Nonmajors and Majors-Level Biology Courses: Student Opinion, Learning and Long-term Retention of Course Material, VOL. 7 LIFE SCIENCES EDUC. 146-154 (2008)(discussing the impact of using "clickers" on undergraduate biology students' learning); John A. Grete & Michael Green, Improving Undergraduate Learning with Computer-Assisted Assessment, 33 J. RES. ON COMPUTING EDUC. 46 (2000) (looking at whether repeated practice with computerized practice tests impacted students' graded in-class exam performance of college education department students); see also Paul T. Wangerin, Action Research in Legal Education, 33 WILLAMETTE L. REV. 383, 385 n.7 (1997) (listing studies evaluating critical thinking and writing skills).

64. Neumann & Krieger, supra note 57, at 386-87 (noting the need to educate colleagues about the fact that empirical research builds upon the work of others and does not need to answer big questions or answer questions conclusively).

65. MACNEALY, supra note 46, at 40-41; see also, Neumann & Krieger, supra note 57, at 354-83 (discussing important components of good empirical research and common errors in empirical research and data interpretation).
empirical research methods will be used. Below is a very brief overview of various study design methods\(^6\) that may be used to empirically examine assessments.\(^7\)

One empirical research method used to examine the impact of a change in teaching or assessment methods involves giving subjects different treatments with as many variables controlled as possible to examine the impact of the particular treatment.\(^8\) This method, called an experiment or quasi-experiment,\(^9\) is discussed at more length in the following two subsections.

Another empirical research method involves the use of surveys. Surveys can be used in many different ways in assessment research.\(^10\) For example, professors could be surveyed to determine the content validity of assessments,\(^11\) practitioners could be surveyed to determine whether existing assessments are authentic and how to improve authenticity,\(^12\) students could be surveyed to determine if the assessment helped solidify their learning, and student surveys also could be used to

---

\(^{66}\) This is not an exhaustive list. For example, there are also ethnographies, quantitative descriptive studies, and prediction and classification studies. For a description of those, see generally JANICE M. LAUER & J. WILLIAM ASHER, COMPOSITION RESEARCH/EMPIRICAL DESIGNS (1988); see also Robert C. Calfee and Marilyn Chambliss, \textit{The Design of Empirical Research}, in METHODS OF RESEARCH ON TEACHING THE ENGLISH LANGUAGE ARTS 43-70 (James Flood et. Al., eds., 2005) (discussing the design of empirical studies); Neumann & Krieger, \textit{supra} note 57, at 353-54 (discussing numerous methods of empirical research).

\(^{67}\) These methods also are useful in examining the impact of a change in teaching methodology on student learning outcomes.

\(^{68}\) For an example of this kind of study, see Curcio et al., \textit{supra} note 61 (testing the impact of multiple practice writing exercises by giving the exercises to one section of a first year civil procedure class and then giving identical final essay exam questions to both the section that had earlier completed the practice writing exercises and to another civil procedure section that had not completed the practice writing exercises, and then comparing the raw score essay exam question point totals of students in the two classes).

\(^{69}\) MACNEALY, \textit{supra} note 46, at 81 (noting that the main difference is that a true experimental design controls for disparity in subjects via random sampling). For a more in-depth discussion of both these methods and their similarities and differences, see LAUER & ASHER, \textit{supra} note 66, at 152-203, and LODICO, \textit{supra} note 47, at 177-207.

\(^{70}\) For a more detailed discussion of developing and analyzing surveys see CRESWELL, \textit{supra} note 47, at 145-54; LAUER & ASHER, \textit{supra} note 66, at 54-81; LODICO, \textit{supra} note 47, at 139-176; MACNEALY, \textit{supra} note 46, at 148-75.

\(^{71}\) For an example of a study that included such a survey, see DIANA DOLMANS ET AL., ASSESSING TEST VALIDITY THROUGH THE USE OF TEACHERS' JUDGMENTS (1992), \textit{available at} \textit{www.eric.ed.gov/} (search "Search Eric Collection" for "Assessing Test Validity Through the Use of Teachers' Judgments"; then follow "Eric Full Text" hyperlink under first result).

\(^{72}\) See discussion \textit{infra} Part IV.C. (discussing ways to determine whether law school assessments measure skills as they are used in practice, thus checking the authenticity of the assessments).
help modify and improve assessments and assessment rubrics.\textsuperscript{73}

In addition to surveys and experimental/quasi-experiments, one can develop empirical studies using discourse or text analysis. This research method examines and categorizes selected portions of written or oral performance.\textsuperscript{74} For example, student journaling could be studied to determine whether integrating skills and professional identity issues into legal education makes a difference in student engagement and satisfaction levels. Finally, after numerous studies have been done in a particular area, a researcher can conduct a meta-analysis—the compiling and re-analysis of similar studies—in order to elucidate new data or establish effective trends.\textsuperscript{75}

Regardless of the research study method, universities, in accordance with federal law,\textsuperscript{76} require researchers to get prior approval of any study using human subjects. This is to ensure that study participants will not be subjected to either physical or psychological harm,\textsuperscript{77} and to ensure that the study will not "deprive a particular class or group of students of a treatment they are entitled to—such as instruction in some skill that one might reasonably expect students to have learned in that class."\textsuperscript{78} Thus, after designing a study, but before implementing it, researchers must obtain the approval of their university's Institutional Review Board.

\begin{itemize}
\item \textsuperscript{73} For an example of a study that used student surveys to gauge whether the use of clickers improved student learning, see Crossgrove & Caron, \textit{supra} note 63.
\item \textsuperscript{74} \textit{MACNEALY, supra} note 46, at 123-47 (discussing discourse analysis). For an example of discourse analysis, see Elizabeth Mertz, \textit{Inside the Law School Classroom: Toward a New Legal Realist Pedagogy}, 60 \textit{VAND. L. REV.} 483 (2007) (examining the linguistic interactions between professors and students in eight different law schools to determine whether law school pedagogy has a shared linguistic structure and epistemological message across diverse classrooms); see also Gay Gellhorn et al., \textit{Law and Language: An Interdisciplinary Study of Client Interviews}, 1 \textit{CLINICAL L. REV.} 245 (1994) (examining the linguistic interactions between professors and students in eight different law schools to determine whether law school pedagogy has a shared linguistic structure and epistemological message across diverse classrooms).
\item \textsuperscript{75} There is, at present, no body of literature on law school assessment work that is ripe for meta-analysis. For a discussion of meta-analysis, see \textit{MACNEALY, supra} note 46, at 109-22. For an example of a study using meta-analysis, see Nancy Flachikov & Judy Goldfinch, \textit{Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks}, 70 \textit{REV. EDUC. RES.} 287 (2000).
\item \textsuperscript{76} For an overview of the statutes and regulations governing the Institutional Review Board process, see Pedro F. Silva-Ruiz, \textit{The Protection of Persons in Medical Research and Cloning of Human Beings}, 46 \textit{AM. J. COMP. L.} 151, 154-57 (1998).
\item \textsuperscript{77} \textit{MACNEALY, supra} note 46, at 94.
\item \textsuperscript{78} \textit{Id.}  
\end{itemize}
C. Issues Arising In Experimental Model Research Design

Although all the methods discussed above can be used to study various aspects of assessment, an in-depth analysis of each method is outside the scope of this Article. Because the experimental/quasi-experimental model can be used for a wide variety of assessment studies, and because it is an area that is riddled with potential research missteps, this subsection focuses on that research method and discusses some key issues to consider when using it.

In planning an experimental/quasi-experimental study, one must decide which variables will be manipulated and which will be held constant. Usually, there is an independent variable that the researcher manipulates, as well as control variables that are identified and accounted for but held constant. The experiment also includes a dependent variable, which may change depending upon how the independent variable is manipulated. Some research models employ this method to measure cause and effect—the researcher manipulates the independent variable to see if it causes a measurable change in the dependent variable.

In developing a study, it is important to consider what one wants to measure, and what will be the independent variable(s). For example, if a professor wants to measure the impact of multiple assessments on a student’s ability to complete a complex factual analysis in an essay exam, the independent variable would be the administration of multiple assessments to the treatment group, while the control group is taught in a different way. One issue confronting professors is whether to have a control group and, if so, how to establish one. There are many different ways to have a control group. For example, one could compare exam answers from prior years when the course was taught using a different method to answers to the same exam questions when the course was taught with innovations. Alternatively, a professor could teach two sections of the same course using different teaching or assessment methods during the semester while administering the same final exam to both course sections. For an example of this method, see Stephen J. Shapiro, The Use and Effectiveness of Various Learning Materials in an Evidence Class, 46 J. LEGAL EDUC. 101 (1996). Similarly, two professors teaching different sections of the same course could collaborate, with one professor adding a specific treatment to her section, while the other taught in the more traditional way. Both sections could be given the same exam questions, and the students’ answers could be compared. For an example of this method, see Curcio et al., supra note 61 (describing impact of practice writing exercises with feedback on student
the more traditional way and does not receive multiple assessments. The dependent variable would be the student’s skill in breaking a rule down into component parts and performing an in-depth factual analysis, as measured by a complex rubric allocating points for each of those factors.

Researchers may have to consider multiple variables in their analysis. In fact, researchers should consider planning a study that involves a multi-variable analysis, called factorial design, since it is virtually impossible to isolate one variable and prove cause and effect. In planning a study, it is wise to hypothesize and plan for these potential multivariate relationships as much as possible. A researcher must also remember that in an educational setting, one cannot conduct the same kind of “pure” scientific study as in a laboratory because the primary goal is to teach students. Thus, there will be many variables that cannot be controlled.

Although it is impossible to conduct an unassailable scientific study in a classroom setting, researchers must still work to ensure that the study is valid and reliable. In experimental empirical research, a

---

85. For a thorough explanation of developing a multi-factor study, see CALFEE & CHAMBLISS, supra note 66, at 61-65.

86. For an excellent explanation of the factorial design process, see MACNEALY, supra note 46 at 88-91. For an example of a study using factorial design, see Barton et al., supra note 33, at 20-26 (2006).

87. “If you ignore factors that influence performance, variability from these sources does not disappear; instead, it confuses the picture. In a well-controlled study, the researcher pins down important sources of variability, to ensure that systematic effects stand out clearly against background noise.” CALFEE & CHAMBLISS, supra note 66, at 58.

88. Id.

89. See infra text accompanying notes 97-100 (discussing variables that may not be able to be controlled, and how to handle those issues).

90. A valid study is one that measures what it is intended to measure. MACNEALY, supra note 46 at 55.

91. Reliability means that others duplicating the same study should get similar results.
fundamental reliability and validity issue is the need to ensure that the control group and the group receiving the treatment are similar.\(^9\) If possible, random assignment of participants to the control or treatment group is the best way to ensure validity and reliability because random sampling helps rule out other explanations for any observed differences.\(^9\)

In a law school study, depending upon the nature of the study, one way to show randomization is to compare LSAT scores, UGPA, or law school GPAs between the control and experimental group.\(^9\) If it is impossible to randomize and if the study has non-equivalent groups, it might be worthwhile to conduct a pre-test to show that the groups are essentially the same in terms of the areas being studied.\(^9\)

Other validity issues include controlling for variables that may affect student performance and outcome,\(^9\) such as teacher experience, events occurring in other classes,\(^9\) data collection at different points in time,\(^9\) subjects discussing the experiment with others who have not yet had the treatment,\(^9\) and subjects dropping out of the study and thus skewing the results.\(^9\) Researchers must be aware of potential variables like these that might impact the study's outcome and need to control for these variables whenever possible. To the extent that the variable cannot be controlled, it is critical that the analysis acknowledge the variable's potential impact on the study's results.\(^10\)

It is important to be aware of the above issues and to address them, as much as possible, before beginning an experimental model. Of

\(^{92}\) Id. at 53.

\(^{93}\) Id. at 70.

\(^{94}\) Id.

\(^{95}\) See, e.g., Curcio et al., supra note 61 (using students LSAT scores and UGPAs to demonstrate that both the control and treatment group were similar).

\(^{96}\) MACNEALY, supra note 46, at 85. For additional suggestions on how to deal with non-randomization and non-equivalent group issues, see id. at 85-87.

\(^{97}\) Id. at 57.

\(^{98}\) Id. at 57; see also CALFEE & CHAMBLISS, supra note 66, at 56. (discussing how there may be conditions occurring over the course of a study, such as participants' personal problems, or a distracting big football game the day of the test, or other events that may affect the study's outcome, and noting that these events should be documented and discussed in the data analysis).

\(^{99}\) MACNEALY, supra note 46, at 60.

\(^{100}\) Id. at 59-60.

\(^{101}\) Neumann & Krieger, supra note 57, at 359 (noting that every study has methodological problems that researchers cannot eliminate, and the key to a good study is to acknowledge those problems responsibly and honestly).
course, other issues may become apparent during the course of the study. Thus, to the extent feasible, a researcher should run a pilot test to see if there are potential problems with the study design in order to remedy as many problems as possible before engaging in the full study.102

D. Analyzing and Interpreting the Research Data

It is beyond the scope of this Article to describe and explain the various methods of statistical measurement and analysis.103 In large part, law professors without statistical training will need to trust their social science or statistician collaborators to do the proper form of statistical analysis. It is important, however, for law professors to have a basic understanding of the statistical methodology, measurement variables, and measures of statistical validity that are being used. Having this knowledge will help to interpret and explain the study’s results, and will make it easier to translate the statistical language into plain English, making the findings comprehensible by non-statisticians.104 It is also helpful to be aware of some frequent potential issues that arise in data analysis and interpretation. Some of those are discussed below.105

Perhaps one of the biggest data analysis issues in assessment research studies involves rater reliability—i.e., is the person scoring the exam applying the rubric in the same way over the course of grading?106 One way to improve reliability is to use multiple raters.107 This may, however, lead to problems with inter-rater reliability—i.e., do different

102. Barton et al., supra note 33, at 33-38 (discussing how they refined their assessment instrument after a pilot study).
103. For a helpful basic description of quantitative statistical analytical measurement terms and concepts, see MACNEALY, supra note 46, at 104-08; LODICO ET AL, supra note 47, at 242-62.
104. Neumann & Krieger, supra note 57, at 390 (noting that researchers with only a social science background “often produce scholarship that cannot be understood outside their social science or that suffers from naivete about law and its institutions”).
105. For a thorough discussion of other potential errors, see Neumann & Krieger, supra note 57, at 368-83; see also CALFEE & CHAMBLISS, supra note 66, at 54-56.
107. Lawrence M. Rudner, Reducing Errors Due to the Use of Judges, 3 PRACTICAL ASSESSMENT, RESEARCH & EVALUATION 3 (1992), available at http://pareonline.net/getvn.asp?v=3&n=3 (noting that multiple raters can improve reliability and discussing ways to increase reliability between multiple raters).
raters using the same rubric get similar results? To minimize this risk, it is important to train multiple assessors in the application of a rubric, and to periodically ensure that the assessors are applying the rubric in the same manner and reaching the same general results.

A common interpretation and analysis problem involves finding significance when there is none or, conversely, failing to recognize a significant finding. This is best addressed through applying the same level of intense analysis and scrutiny to the empirical data and analysis as one applies to other scholarly endeavors. For interpretation issues, especially finding no statistically significant differences when one hypothesized that such differences would likely exist, a potential remedy is to think about dividing the data in a different way. If, after doing this, there still is no statistically significant finding, it does not mean that the study was useless. In fact, disproving a hypothesis may be a very useful undertaking. The key is to admit that the research results did not confirm the original hypothesis and then attempt to explain potential reasons for the surprising outcome.

Researchers also should avoid the temptation to overstate the significance of a particular finding or use the methodology to prove,
rather than explore, a particular point or point of view. In social science research, the end of one study often is only the beginning of the exploration process. Thus, one should not try to do too much with any one study and should recognize that, sometimes, simply designing a cost-effective duplicable model is a contribution to the literature.

Most importantly, a researcher must acknowledge all weaknesses in the interpretation of the data, including uncontrolled variables and the study's limitations.

Finally, as Professors Neumann and Krieger note, social science empirical research standards and methodology is foreign to many law professors and, thus, law professors incorrectly may believe that a study lacks value if it does not definitively answer a question, it duplicates the results of others, it makes small steps in proving a point, or it fails to prove the point altogether. Until the legal academy becomes better acquainted with what empirical work can and cannot do, legal scholars must educate the reader regarding these issues so that their work is not discounted by those in the academy who are unfamiliar with the incremental work involved in empirical research studies.

IV. APPLYING THEORY TO PRACTICE—EMPIRICAL RESEARCH OF LAW SCHOOL TEACHING METHODOLOGIES AND ASSESSMENTS

Regardless of whether law professors believe the status quo is an effective way to assess students, or whether they believe that assessments should change, one area of legal scholarship that lags far behind other disciplines is the systematic study of teaching and assessment, and its impact on student learning. Studying assessments may have many different purposes. One purpose may be to demonstrate the strengths and weaknesses of existing assessments and to improve that further study may be warranted.

114. See id. at 368-75 (discussing the problems with over-generalizing and using methodology to prove, rather than explore, a particular point).
115. CALFE & CHAMBLISS, supra note 66, at 76 (noting that "[a] series of modest but well-designed studies is likely to be more informative than a single humongous effort").
116. See Curcio et al., supra note 61, at 312-13 (discussing how the authors believed that one contribution made by their study concerning the impact of practice writing exercises on student essay exam performance was the study's duplicable design).
118. Id. at 386-88.
119. For an example of what other disciplines, in particular medicine and business, are doing in terms of assessment research, see supra note 62.
upon those assessment methods.\footnote{120} Another purpose may be to develop a duplicable study model that can be used to test whether different assessments or teaching methodologies result in improved student learning.\footnote{121} Alternatively, one may choose to study assessments to demonstrate that it is possible to validly, reliably and cost-effectively assess a wider range of skills, thus prompting more professors (and perhaps bar examiners) to incorporate these kinds of assessments into their tests.\footnote{122}

A. Basic Assessment Issues: Assessments Must be Valid and Reliable

One key area of assessment research involves examining whether assessments are valid and reliable. Validity involves ensuring that the assessment tests the knowledge and skills that it is designed to measure.\footnote{123} The main concerns of law school exams generally are construct validity and instructional validity.\footnote{124} Construct validity involves demonstrating that the assessments produce "learner behaviors that bear a direct link to the cognitive activity you want to assess."\footnote{125} For example, if a professor seeks to assess whether a student can "think like a lawyer" (i.e., synthesize legal rules and critically analyze and apply them in a given fact situation), the exam should actually assess these skills.\footnote{126} Construct validity also encompasses the issue of whether

\footnote{120. See, e.g., Lynn M. Daggett, \textit{All of the Above: Computerized Exam Scoring of Multiple Choice Items Helps to: (A) Show How Exams Worked Technically, (B) Maximize Exam Fairness, (C) Justly Assign Letter Grades, and (D) Provide Feedback on Student Learning}, 57 J. LEGAL EDUC. 391 (2007).}

\footnote{121. See, e.g., Curcio et al., supra note 61, at 271, 290.}

\footnote{122. Barton and her colleagues gave a detailed explanation of the steps that they took to ensure the reliability and validity of the Standardized Client model they used. Barton, et. al, supra note 33, at 33-41. This Standardized Client model is now being used as part of the Franklin Pierce Daniel Webster Scholar Program, an alternative lawyer licensing model. See Garvey & Zinkin, supra note 34, at 121-122. John Garvey, Director of the Daniel Webster Scholar Program, in a discussion with the author, indicated that the Standardized Client Model will also be used to compare client interviewing skills of recently licensed graduates of the program with recently licensed lawyers who did not complete the program.}

\footnote{123. See Daggett, supra note120, at 393.}

\footnote{124. For an in-depth explanation of these, and other, validity concepts, see generally \textit{TEST VALIDITY} (Howard Wainer & Henry I. Braun eds., 1988).}

\footnote{125. \textsc{Martín L. Tombaři & Gary D. Borich}, \textsc{Authentic Assessment in the Classroom: Applications and Practice} 53(1999) (discussing the basic concept of construct validity and methods to assure assessments have construct validity).}

\footnote{126. Daggett, supra note 120, at 394 (noting that an element of construct validity with law school exams is their goal of measuring whether students can "think like a lawyer").}
there is a defensible rationale for the scoring rubric.\textsuperscript{127}

Instructional or content validity looks at whether the assessment accurately reflects the goals and objectives of what has been taught and emphasized in the classroom.\textsuperscript{128} One law professor explains content validity in this way: "[D]oes a torts exam, for example, measure the torts concepts (such as the concept of tortious intent) and skills (such as being able to spot viable legal claims in a fact pattern and predict their likely outcome in court) students are expected to have learned in the course."\textsuperscript{129}

In addition to being valid, assessments must also be reliable because "[e]xams that lack reliability produce scores that are due significantly to chance and other irrelevant factors, rather than scores that are due to different levels of student performance, and are thus not a good basis for assigning letter grades, or any other purpose."\textsuperscript{130} Law professors should be concerned with three reliability issues: internal consistency, inter-scorer reliability, and intra-scorer reliability.\textsuperscript{131} Internal consistency means that questions assessing the same skill should result in similar quality of answers; inter-scorer reliability refers to the idea that different scorers should arrive at similar marks for the same answers; and, intra-scorer reliability looks at whether the marks given by the same teacher at different times remain constant.\textsuperscript{132}

Both validity and reliability issues provide fertile ground for empirical assessment research. The following sections suggest numerous different studies that can be developed to explore reliability and validity issues for existing law school assessment methods as well as for some of the alternative methods explored in Part II.

\begin{itemize}
\item \textsuperscript{127} John Crawford & James C. Impara, Critical Issues, Current Trends and Possible Futures in Quantitative Methods, in HANDBOOK OF RESEARCH ON TEACHING 133, 143 (Virginia Richardson ed., 4th ed. 2001) ("[T]he scoring criteria and rubrics should not be constructed capriciously, just because others have used similar schemes. One should have a defensible rationale for whatever scoring process is used.").
\item \textsuperscript{128} TOMBARI & BORICH, supra note 125, at 56-59 (discussing instructional validity and how to ensure assessments are instructionally valid).
\item \textsuperscript{129} Daggett, supra note 120, at 394.
\item \textsuperscript{130} Id. at 395.
\item \textsuperscript{131} Maziam Alias, Assessment of Learning Outcomes; Validity and Reliability of Classroom Tests, 4 WORLD TRANSACTIONS ON ENG'G & TECH. EDUC. 235 (2005).
\item \textsuperscript{132} Id. For an explanation of how to measure reliability by comparing or correlating students' scores on an exam, see Daggett, supra note 120, at 396.
\end{itemize}
B. Studies Involving Design and Development of Rubrics

Assessment research design may require that the researcher develop an assessment rubric that identifies the learning outcome being assessed and the criteria for assessing each outcome. How does one develop a valid and reliable rubric? Professors Neumann and Krieger provide the following insight into the development of standards of measurement:

[Valid standards of measurement cannot be created out of whole cloth by individual researchers but must be based on prior qualitative and quantitative research on similar issues, impressionistic evidence in the field and theoretical models. Valid measurement standards, therefore, develop as a product of a social enterprise of scientists building on each other’s work. While precise measurement of certain qualitative variables may be impossible, empirical researchers consider these ongoing efforts as a way of continuously developing more accurate standards.]

Thus, one area ripe for study is the development of valid and reliable rubrics for both existing and alternative assessments. Designing valid and reliable rubrics would be a significant contribution to assessment research because these rubrics would provide...
law professors with grading guides and thus, possibly pave the way for others to employ similar assessment methods. Additionally, a well-designed rubric is often the starting point for assessment research.138

C. Empirical Studies Examining Construct Validity

Another area of research involves studying assessments for construct validity.139 For example, one could look at whether an assessment is authentic—does it correspond to how the doctrine and skills are used in practice?140 A study could examine whether concordance exists between how law professors assess students’ skills and how those skills are evaluated or used in practice.141 Alternatively, one could look at whether the material being taught and assessed in particular doctrinal courses had content validity in terms of a particular area of practice. For example, a potential study could involve a survey of products liability lawyers to see what issues most often arise in their practice, followed by a content analysis of various products liability exams to determine whether the material being tested aligns with the issues that arise most frequently in practice. One could also simply study whether law school exams, in general, assess what professors teach.142 Another construct validity study could build upon an initial

138. See id. (using an empirically designed and tested rubric as the basis for their legal writing study).

139. For a description of construct validity, see supra text accompanying notes 125-127.

140. “One theme running through the many contemporary versions of experiential learning is that of ‘authenticity’—the correspondence, in some way or other, of learning to the world of practice that exists outside of teaching institutions.” Karen Barton et al., Authentic Fictions: Simulations, Professionalism and Legal Learning, 14 CLINICAL L. REV. 143, 145 (2007). For a study discussing reliability and validity issues arising in authentic assessments, see Raimundo Olfo & Hildaura Zulantay, Reliability and Validity of Authentic Assessment in a Web Based Course, 10 EDUC. TECH. & SOC’Y 156 (2007).

141. See, e.g., Akshat Tewary, Legal Ethics as a Means to Address The Problem of Elite Law Firm Non-Diversity, 12 ASIAN L.J. 1, 11 (2005) (arguing that law school grades have poor predictive value in terms of effective lawyering and citing a study that shows no correlation between law firm partner income and law school grades or law review membership); David B. Wilkins & G. Mitu Gulati, Why are There so Few Black Lawyers in Corporate Law Firms? An Institutional Analysis, 84 CAL. L. REV. 493, 526-27 (1996) (discussing how law school grades do not reliably reflect lawyering skills). One testable hypothesis is that law school exams teach students ineffective lawyering skills because students get credit for anything that is correct and there is little, if no penalty, for including a fair amount of irrelevant information. One interesting study would be to ask judges and practitioners to assess the quality of both in-class and take-home exams and see how their assessments correlate to the students’ course grade.

142. For an example of a construct validity study in which the researcher sought to measure whether the exams actually assessed what was taught, see DOLMANS ET AL., supra
study by Professor William Henderson, in which he suggests that law professors may be unaware of, or discount, the unintentional role of test-taking speed as an independent variable in in-class timed exams.\(^{143}\) One could examine both whether test-taking speed itself had content validity in a given set of doctrinal courses, as well as the role it plays in students' performance on law school exams.

**D. Empirical Studies Examining Reliability**

Assessment studies could also be developed to test whether law professors' grading is, on average, reliable—an area ripe for study. In fact, so little work has been done in this area that recent scholars are quoting a 1924 study to support their arguments.\(^{144}\) Reliability studies could be conducted in numerous ways. For example, a researcher could examine individual student's raw scores on different questions testing similar skills or material for correlations.\(^{145}\) Or, to study intra-rater reliability, a professor could grade all exams and then re-grade all of them a week later using the same rubric.\(^{146}\) Inter-rater consistency could be examined by two professors, each grading the other's' exams using the other's' rubrics.\(^{147}\) Or, multiple professors could be asked to grade

---

note 71.

143. *See* William D. Henderson, *The LSAT, Law School Exams and Meritocracy: The Surprising and Undertheorized Role of Test-Taking Speed*, 82 TEX. L. REV. 975 (2004) (finding that test-taking speed was an independent variable in law school exams and suggesting that follow up studies on this issue, and on the issue of whether timed law school exams have a disparate impact on one or more ethnic subgroups were areas ripe for further exploration.); *see also*, Ruth Colker, *Extra Time As An Accommodation*, 69 U. PITT. L. REV. 413 (2008) (performing a small empirical study examining the correlation between student performance on in-class timed exams and take home exams).

144. The 1924 study was conducted by Professor Benjamin D. Wood to measure the reliability of law school essay and multiple choice exams. The study, which could be duplicated ninety years later, has been cited recently by at least two legal scholars. *See, e.g.*, Colker, *supra* note 143, at 457; Linda R. Crane, *Grading Law School Examinations: Making a Case for Objective Exams to Cure What Ails “Objectified” Exams*, 34 NEW ENG. L. REV. 785, 797-98 (2000).


146. This is an example of intra-rater reliability—something that other studies have shown is often low. *See, e.g.*, MUNRO, *supra* note 49, at 108-09 (discussing a study showing low intra-rater reliability even in a system that had more techniques for ensuring reliability than the average law professor uses in grading exams).

147. This is also likely to show low reliability based on other preliminary studies, such as the one cited by Professor Munro, in which a California bar exam candidate had only a sixty-seven percent chance that two examiners would agree as to whether the answer passed or failed. *Id.* at 108.
the same exam without an agreed-upon rubric to determine if there is reliability between professors in terms of what constitutes a good exam answer.\footnote{148}

Finally, one could test whether law school exams purporting to test the same basic cognitive skill set are reliable even as to that skill set. For example, some may claim that law school assessments are de facto reliable because students generally get around the same grades in all their courses, especially in the first year. One could examine whether there is a statistically significant correlation between students' grades in multiple first-year classes, and whether the statistical significance exists across the board or exists most strongly for those who are doing extremely well or extremely poorly.

If studies show an expected lack of reliability in current assessment practices,\footnote{149} the next area for study would be how to increase reliability given the high stakes involved in law school assessments and grades. Studies could look at such issues as how best to develop rubrics to ensure as much reliability as possible.\footnote{150} Studies also could experiment with different grading methods, such as looking at whether periodic “grading exchanges” between professors using an agreed-upon rubric increases reliability,\footnote{151} or whether the use of lists, rating guides, and model responses makes a difference in reliability.\footnote{152} These reliability studies could be done with existing assessments, as well as assessments measuring skills not commonly assessed in doctrinal courses, such as listening, factual investigation, client interviewing and counseling, and the ability to see the world through the eyes of others.\footnote{153}

\section*{E. Studies Examining the Impact of Various Assessment Formats}

Formative assessments are an educational tool used to inform

\footnote{148. A similar study was conducted in 1968 and, not surprisingly, found a lack of a coherent grading standard. Friedland, \textit{supra} note 19, at 184 (discussing a study performed by Stephen Klein and Frederick M. Hart).}

\footnote{149. See \textit{id.} at 183-86 (arguing that law school essay exam grades are unreliable); Daggett, \textit{supra} note 120 (suggesting that multiple choice quizzes given throughout the semester and in conjunction with a final exam are a good way to counter some of the reliability issues raised by essay exams and a single end-of-semester exam).}

\footnote{150. See \textit{supra} Part IV.B. (discussing development of grading rubrics).}

\footnote{151. This idea was taken from Early, \textit{supra} note 109, at 8.}

\footnote{152. See Schwartz, \textit{Teaching Law By Design}, \textit{supra} note 106, at 407 (noting that studies in other disciplines have found that these strategies improve reliability).}

\footnote{153. For suggestions on how to assess these kinds of skills, see \textit{supra} Part II.}
students about their strengths and weaknesses. Summative assessments are used to grade students. Of course, summative assessments can be formative. Education theorists assert that more frequent law student assessments, both of a formative and summative nature, will improve student learning. This proposition can be tested. One way to do that is to compare student final exam performance between students who got multiple assessments and those who did not. Another study might involve examining whether the use of multiple assessments (or multiple types of assessments), as opposed to a single end-of-semester assessment, leads to a compression of the raw point score spread, which may, in turn, lead one to question the validity of a single exam grade. Research studies could also seek to determine the impact of more frequent assessments on students’ mental health and well-being—i.e., do the more frequent assessments lessen or increase students’ stress level and engagement or disengagement?

154. STUCKEY ET AL., supra note 2, at 255. Ungraded homework assignments, ungraded quizzes and using “clickers”—a process by which students are asked to “click” their answers to hypothetical problems and the results are displayed on a bar graph—are all ways to engage in formative assessments and all are ripe for study as to whether adding formative assessments to a class improves student learning. For more information about the use of clickers in law school classrooms, see Nancy G. Maxwell, From Facebook to Folsom Prison Blues: How Banning Laptops in the Classroom Made Me a Better Law School Teacher, 14 RICH. J.L. & TECH. 4, 16-17 (2007). For information on how to develop a study looking at the impact of clickers on law student learning, see Crossgrove & Caron, supra note 63.

155. See STUCKEY ET AL., supra note 2, at 255.

156. For example, professors can give quizzes throughout the semester. See, e.g., Daggett, supra note 120 (discussing the value of multiple choice quizzes in large-section law school courses).

157. See, e.g., STUCKEY ET AL., supra note 2, at 255-261; Friedland supra note 19, at 188; Munro, supra note 49, at 74-75.

158. See Curcio et al., supra note 61, at 300-302 (finding that multiple practice essays only had a statistically significant impact on students with above-the-median LSAT scores and UGPAs). One interesting study would be to build upon the work done by Professor Daggett. She believed that giving multiple choice quizzes throughout the semester served a formative evaluative purpose. Daggett, supra note 120, at 413-14. It would be interesting to give the same final exam to students getting multiple choice quizzes and those in a different section of the same course who did not get the quizzes to see if the quizzes improved students’ final exam performance. A slight twist would be to perform a study comparing students who had multiple assessments with professorial feedback with those who had multiple assessments with a rubric but only student self-reflection to see if the professorial feedback made a difference in student learning outcomes.

159. See Morris, supra note 9, at 449-453 (discussing the impact of first year grades on students’ mental health and their engagement in the learning process).
F. Studies Examining the Impact of Assessment Format and Content on Law Student Diversity

Finally, assessment research in many of the areas suggested above could also examine the impact of different types of assessment on learning outcomes and exam performance of students with economically-disadvantaged or racially and ethnically diverse backgrounds.\textsuperscript{160} Does assessing a wider range of skills, or assessing differently, change the exam performance differential between students of color and white students? A preliminary study by Professor William Henderson suggests that performance gaps between white students and students of color are narrowed when students are given take-home exams and papers rather than timed in-class exams.\textsuperscript{161} Professor Henderson urged further study of this issue using a much larger sample size.\textsuperscript{162}

G. Generating Ideas

The above are simply a preliminary list of potential areas of assessment research study. These suggestions are meant to stimulate thought about how to teach and assess different skills and how to begin an empirical examination of existing and alternative assessments. So little work has been done in this research area that almost any empirical study a law professor does to examine assessment will be a valuable addition to the academy's knowledge about teaching and assessment. Studies do not have to be large, complicated multi-year undertakings; small studies that build a model for further research, or that suggest other areas to be explored, can be extremely valuable to others seeking to engage in this kind of work.\textsuperscript{163}

V. CONCLUSION

Existing law school assessments have proven inadequate to measure the majority of skills necessary for effective lawyering.

\textsuperscript{160} See, e.g., Ian Weinstein, Testing Multiple Intelligences: Comparing Evaluation by Simulation and Written Exam, 8 CLINICAL L. REV. 247 (2001) (examining whether evaluating different skills eliminated the disparities in grades between white students and students of color).

\textsuperscript{161} Henderson, supra note 143, at 983.

\textsuperscript{162} Id. at 1039.

\textsuperscript{163} CHAMBLISS & CALFEE, supra note 66, at 76.
Additionally, current and future law students may be caught in an assessment system that inaccurately and unfairly curtails their employment opportunities,\textsuperscript{164} affects their sense of self-worth,\textsuperscript{165} and stands as an unjustifiable barrier to entry into the profession for people of color,\textsuperscript{166} simply because law professors continue to do things the way they have been done for decades.

The legal academy should not accept the status quo as the best available option unless, and until, it devotes serious time and energy to exploring alternatives or proving that the existing assessment methods are the most valid, reliable, and cost-effective methods currently available to assess law students and prepare them to be good lawyers.

Whether alternative assessment methods will improve student learning, make law students more effective lawyers, have any impact upon diversity in law school and in the bench and bar, or lead to less student disengagement and disillusionment is, as of yet, unknown. Given the high stakes of the law school assessment system, however, the legal academy owes law students a serious examination of these questions. One way to begin that examination is through empirical work.

Empirical examination of assessment methods, especially when done in collaboration with a social scientist, is not beyond the reach of any law professor. It simply requires the same level of deep critical thinking that law professors employ in other scholarly pursuits and a desire to explore issues that may make a difference in the lives of students and, ultimately, the profession.

---

\textsuperscript{164} See supra note 8 and accompanying text.
\textsuperscript{165} See supra note 9 and accompanying text.
\textsuperscript{166} See supra notes 5-7 and accompanying text.