Jeffrey L. Bernstein, EASTERN MICHIGAN UNIVERSITY, jeffrey.bernstein@emich.edu

Unifying SoTL Methodology: Internal and External Validity

ABSTRACT

A broad consensus exists that the use of appropriate methods are important in the Scholarship of Teaching and Learning. However, methodological controversies arise around what constitutes acceptable evidence, if one needs a control group, how generalizable results must be, and other similar issues. Much SoTL work, I argue, asks questions about how much a particular treatment (innovation) caused an effect (student learning), and how the results found in one particular context can be extended outside that context (generalizability). These concepts, known as internal validity and external validity, respectively, provide a common point of departure for much scholarship on teaching and learning. This paper addresses these concepts and demonstrates how they can unite much of what divides us within the methodological realm of SoTL.

KEYWORDS

SoTL, methodology, internal validity, external validity, generalizability

My formal introduction to the Scholarship of Teaching and Learning came as a Carnegie Scholar with the Carnegie Foundation for the Advancement of Teaching in 2005-2006. As I gathered with some brilliant minds in teaching and learning—talk about the imposter syndrome!—I was in awe of the work some of my colleagues were doing. Compared to them, I was an amateur, a poser. Yet as we began to move into our teaching projects and to discuss how we would collect and analyze our data in an attempt to go public with our results, I suddenly became one of the most sought-after people in the room. I was, you see, a "methods person." Trained as a quantitative social scientist with experience in research design, survey construction, and statistical analysis, I could help my colleagues to gather "better" data (as some of us might have perceived) for their projects, and to analyze the data they had collected.

As our Carnegie residencies continued, I realized I was taking at least as much out of the methods "bank" as I was putting in. Colleagues taught me methodologies for "close reading" of student essays, which I was able to apply to student papers (see Bass & Linkon, 2008, for a useful discussion of this approach). While my approach at first was quantitative (scoring essays based on an evaluation rubric, and then testing to see if scores improved by statistically significant amounts as the term proceeded), I eventually found value (and comfort) in using the writing itself, rather than my scoring of it, as evidence. The narrative, rather than the statistic, became a valuable tool in my arsenal for demonstrating the impact of classroom innovation. Pat Hutchings's commentary on my data analysis during one of my presentations stays with me: "I found your statistical evidence to be convincing, but I found your stories to be compelling."

In our residencies, my Carnegie cohort and I established a true methodological "trading zone" (Huber & Hutchings, 2005), sharing what we knew of our methodological backgrounds and offering it up to others who were asking questions for which our techniques could be useful. Moreover, as we looked at the exciting things our colleagues were doing with their own projects, we began to imagine how *their* methods could enhance *our* inquiries. Our personal relationships with one another facilitated such trading, as did our commitment to doing good scholarly work. These helped us overcome the tensions that attend to methodological disputes in any field.[1]

I write this essay from a position of privilege if, as Chick (2013) and Grauerholz and Main (2013) convincingly argue, there is a bias toward social science methodologies within SoTL. That some members of my cohort perceived my methods as "better" (an opinion I did nothing to change at the time further suggests that this might be the case. The ability to use tests of statistical significance to demonstrate a point seems to carry more cache than the ability to document student work based on a close reading of an assignment. When I argue that we need to broaden our definitions of what makes for good evidence in the Scholarship of Teaching and Learning, I may have less at stake than others do; my work, after all, fits within the dominant paradigm. Nevertheless, as someone who practices this "acceptable" form of research in the Scholarship of Teaching and Learning, I may be best positioned to broaden what this dominant paradigm means and perhaps, in a small way, to subvert it.

I begin this essay with a discussion of methods within the Scholarship of Teaching and Learning, borrowing language from Chick (2013) to agree that while there is indeed a broad consensus that methods matter, the "Big Tent" she desires is not as robust as it ideally should be. I then propose the ideas of *internal validity* (the degree of confidence we have that our treatment causes the effect) and *external validity* (the degree of confidence we have that our results are generalizable to the big world out there) as a unifying framework for SoTL methodology, and discuss each of these concepts in turn. I conclude with remarks on the future of methodological issues in our work and with a plea for openness and inclusion in our methodological orientations.

METHODS AND THE SCHOLARSHIP OF TEACHING AND LEARNING

From the early days of the Scholarship of Teaching and Learning movement, concerns about methodology have been part of the conversation. Following Boyer's (1990) seminal *Scholarship Reconsidered,* which introduced the term "scholarship of teaching," Glassick, Huber, and Maeroff (1997, p. 25) began to operationalize what would need to occur were the study of teaching to be treated as scholarship. Their book suggested that effective scholarly work, no matter the context, exhibits clear goals, adequate preparation, *appropriate methods,* significant results, effective presentation, and reflective critique (emphasis mine). They note, presaging the debate within the Scholarship of Teaching and Learning community, that "the choice of method is critical because upon it depends not only the project's chances for success…but also the likelihood that colleagues will understand and accept the project. Scholars who favor quantitative studies, for example, may be reluctant to accept findings based on qualitative approaches, whatever the intrinsic merit of the work" (p. 28). Glassick et al. suggest that while using the appropriate methods matters a great deal, the definition of "appropriate" may rest in the eye of the beholder, and may influence how the work is received.

Felten's (2013, p. 1230 discussion of principles of good practice in the Scholarship of Teaching and Learning also highlights methodology, arguing that good SoTL is "methodologically sound." He argues that "[r]egardless of the methods employed, good practice in SoTL requires the intentional and rigorous application of research tools that connect the question at the heart of a particular inquiry to student learning" (p. 123). Felten joins Huber and Morreale (2002), who stress that the Scholarship of Teaching and Learning should respect the disciplines, in particular in the way disciplines ask different questions and use different methods to assess pedagogical practice.

Chick (2013) uses the competing metaphors of the Big Tent (in which we welcome everyone) and the Family Table (where some conversations are welcomed and some are shut down) in her discussion of methodology and SoTL. As Glassick et al. (1997), Huber and Morreale (2002), Felten (2013), and others suggest, the scholarship of learning ought to welcome a wide range of methods, deployed in accordance with the particular questions one asks, and the particular audience one addresses. However, Chick argues that certain methods seem to be favored over others: as Gale (2005, p. 5) notes, work in the humanities in general is sometimes characterized as being "academically soft." Maurer (2011) argues for a particular brand of research when he strongly encourages control groups and experimental designs, while Gurung and Schwartz's (2009) book on how to do SoTL emphasizes the use of statistics more than a humanist might otherwise desire. Grauerholz and Main (2013) suggest that fallacies associated with SoTL (such as the need for a control group, or generalizability), can discourage those from the non-preferred disciplines from doing this work; losing such voices would remove valuable insights and energy from the field.[2]

Chick (2006) cites Shulman in arguing that we should not become a family table, with conversation that becomes closed and insular. I would extend the family metaphor in a different way. Often, a family table may be a tense place. Uncle Bob and Aunt Gloria might be nursing a long-term feud. An outsider observer might not understand this. She might listen to each and think that they are reasonable people who have more in common than they perceive; such an observer may see no obvious reason for a long-running, continuing argument. The participants, entrenched in their narratives, may be unable to see past their dispute and realize they are speaking the same language. These characteristics of the family table hamper our attempts to develop a methodological big tent. When we step back, we can see that the big questions we ask about methodology are essentially the same, and our answers may not be as different as we all might think. At the very least, the aims that we have are often very much in common (to improve student learning through effective action inside and outside the classroom), even if the way we address such goals are different.

A SIMPLE PERSPECTIVE ON METHODOLOGY

My point of departure is Pat Hutchings's (2000) typology of questions in the Scholarship of Teaching and Learning. Hutchings identifies four types of questions: (a) what works; (b) what is it/what does it look like; (c) visions of the possible; and (d) theory-building questions/questions *about* SoTL.[3] My work here focuses most directly on the first two questions: "what works?" and "what does it look like?," although there may be applications to the other kinds of question as well. For most of us, I suspect, the "what works?" question brought us to SoTL in the first place. We teach something and know that our students' learning is not all we want it to be. Perhaps our students are struggling to understand evolution (Nelson, 2000), or the nature of proof in mathematics (Bennett & Dewar, 2013), or sustainability leadership (Burns, 2016). We think deeply over the matter, talk to colleagues, read the literature, and devise a new approach to teaching the topic in question. If we are well versed in the Scholarship of Teaching and Learning, we may design an inquiry right from the beginning, gathering data and approaching the question of whether our innovation is working, hoping to go public with our evidence. However, for those who know little about the Scholarship of Teaching and Learning, the desire to improve practice is all that they need. They try out new approaches, gather "data" by informal observations of what they see happening, and draw their conclusions. It may be only later that questions of systematic inquiry enter the equation and they consider doing a more formal inquiry to demonstrate that their approach works, and then go public with their results.[4]

Likewise, some of us who are teaching see something happening in our classes, and want to describe it and explain it. Rather than limiting our question to whether something works, we want to create a rich, thick description, so others can learn from what we are seeing. What does it look like when students struggle through moments of difficulty (Salvatori & Donahue, 2005) ? What is it about effective cognition in history that looks different from weaker approaches to using historical evidence (Wineburg, 2000? When studying this type of question, we might consider Shulman's (2013) advice in his keynote address at the International Society for the Scholarship of Teaching and Learning Conference: rich descriptions of particular interventions, which delve deeply into the context in which learning takes place, are more valuable than more generalized studies, since "generalizations decay."

For those who are working on "what works?" or "what is?" questions, and hoping to go public with the results, two essential questions arise. For a SoTL project rooted in a "what works?" question we first ask if the intervention made a difference in student learning. This applies no matter what the intervention, no matter the desired difference in learning, and no matter how we choose to measure it. If I try to use group discussions to motivate engagement in a sociology class, I must be able to show that this intervention has led to enhanced student learning, in a rich, context-specific way. I need to document, through appropriate means, that what I have done has made a difference. The relationship may be more complicated than a simple "when I do A, students do B"—other factors might have contributing impacts. At the end of the inquiry, however, I must be able to show some impact on what the students have learned. A similar consideration applies to SoTL projects rooted in "what works?" questions. Did the exploration result in a description of some aspect of how students learn that is accurate and relevant? At the end of the inquiry, however, I must be able to show that what I have focused on is relevant to the particular process of learning related to my question and that the description I have developed accurately reflects the evidence that I have gathered.

Second, we need to address the extent to which our work is generalizable to a broader audience. This is not to suggest that all work must be generalizable to all audiences at all times. However, it does suggest that for a scholar's work to be valuable (and to be competitive for publication), it must teach lessons that go beyond the narrow context in which the pedagogical innovation was implemented. This generalizability can come in various forms: my work might be generalizable to yours because we teach in the same discipline, or teach the same broad topics, or teach the same types of students, or use the same pedagogical approaches. Work done in a narrow context, and only applicable to that very narrow context, can be illuminating, but is less valuable to the field as a whole. If a large part of the goal of the Scholarship of Teaching and Learning is to improve teaching practice, and to be a mechanism by which the craft of teaching improves itself (Hutchings & Shulman, 1999), then we must ensure that our work does not just improve our own teaching, but also provides an opportunity for the field to advance.

Borrowing from the social sciences, we can label these two ideas as internal validity (demonstrating that our intervention made a difference, i.e., that the treatment caused the effect) and external validity (demonstrating that our work is generalizable outside our narrow context). As I illustrate below, what gives us internal validity sometimes costs us external validity, and vice versa. Some studies excel at one, while others blend both. Putting aside the operationalization of these ideas for a moment, I would argue that these are the two most essential questions we must ask about any teaching innovation. We need to know if it worked, and we need to know how generalizable it is to a larger context.

THE DRUNK-DRIVING STUDY: INTERNAL AND EXTERNAL VALIDITY

I illustrate an application of internal and external validity with an extended reference to an article about college students' willingness to get into a car with a peer who has had too much to drink. While the piece has nothing to do with SoTL, it illustrates better than any other piece I know my central methodological points about internal and external validity.

When college students face the choice of whether or not to get into a car with a peer who has had too much to drink, the behavior of other peers is important. If friends do not get into the car, the person in question might also refuse. If friends do get into the car, the person in question might do so as well. How can we test or this tendency? A survey is unlikely to work, as people will (a) offer the socially desirable response; and/or (b) be unable to predict what they would do in an intense, albeit hypothetical, scenario. Observing this phenomenon in its natural setting (such as by unobtrusively hanging around outside a bar, or party) would provide useful data, although the logistics (and ethics) of such a study would be challenging.

Powell and Drucker (1997) offer an experimental approach to this question; I define an experiment as a research design wherein the researcher controls who gets the treatment, and under what conditions. Students at the University of Hartford volunteered for this study, in which they were told they would be blindfolded and driven around campus to test their spatial orientations. They were told to report to a location on campus and await the driver, who would do the study with them. Unbeknownst to the participants, the investigators randomly divided them into four groups, each of which received one particular treatment as described below:

- Treatment 1: The driver showed up sober, but holding a fake beer in his hand.
- Treatment 2: The driver showed up "drunk"—slamming on the brakes, slurring his speech slightly, and walking unsteadily. The car was filled with discarded beer cans.
- Treatment 3: Same as treatment 2, except that this time, there was a second person (the confederate, who was part of the research staff) who got into the car with the "drunk" driver.
- Treatment 4: Same as treatment 3, only this time the confederate refused to get into the car.

The respondent would then get into the car (or decide not to do so), whereupon the researcher would pretend to leave for a moment, and then come back, end the study, and debrief with the participant. In total, 10 people were part of each treatment. Of those who received the first treatment, nine out of 10 got into the car. In the second and third treatments, all 10 got into the car, while in the final treatment, only one got into the car.

What makes this study exceptional is its internal validity. If we restrict our prime focus to treatments 3 and 4, the behavior of the confederate seems to cause the behavior of the research subject. Since the choice of who got which treatment was randomly determined, the people given treatment 3 should be, on average, just like those given treatment 4. The two groups were treated in exactly the same way—they did the study at the same period, in the same location, with the same research staff—with one exception: the behavior of the confederate. When the confederate got in the car, the research subject followed—and, with one exception, when the confederate refused, the subject did as well. There is no other credible explanation for differences between these two groups than the behavior of the confederate.

The key weakness of this study, however, is external validity. The study is simply not generalizable to the big world out there. Some of this is easy to address. For example, the research was done at the University of Hartford, which raises questions about whether students at other universities would behave in a similar way. This seems easy to address; it is not too much of a stretch to argue that around issues of alcohol, college students nationwide have more in common than not. What is harder to address, however, are the ways in which this situation feels much more like a laboratory than the real world; it is here that the study runs into trouble.

The students in this experiment faced the decision whether or not to get into the car with a drunk driver. The decision occurred in the daylight, on campus, as part of a university-approved protocol, with a complete stranger planning to drive the car at slow speeds around campus. This context would be different from the setting in which these decisions are typically made. For starters, most people getting into the car with a drunk-driver would be with a friend. It is more likely to be at night, perhaps in an unsafe area. The subject may not be sober—after all, if you were sober, you would take the keys rather than getting into the car with a drunk driver. The subject in our experiment could just turn around and go home; that might be harder if the bar at which one is drinking is a distance from home.

The drunk-driving study provides a rich example of the interplay between internal validity and external validity. When we can control the experimental conditions perfectly, as in this article, we can be confident the treatment is causing the effect (internal validity). However, the more we manipulate nature, the less we can generalize from our study to the larger context (external validity); this is similar to Lewis, Perry, and Murata's (2006, p. 8) observation that "the very qualities that suit an innovation to controlled trial may handicap it at the later stage of broad dissemination." While there is often a tradeoff between the two, as illustrated above, these concepts can become a unifying framework for methodology in the Scholarship of Teaching and Learning. It is to this discussion that I now turn.

INTERNAL VALIDITY: DID OUR WORK MAKE A DIFFERENCE

As scholars of teaching and learning, we aim to become experts at identifying promising pedagogical approaches, modeling them in the classroom, and judging if they work. The development of expertise, according to Tetlock (2006), is driven in large measure by the extent to which the individual predictor is a fox or a hedgehog.[5] Tetlock borrows from the lesson originally noted by the Greek poet Archilochus and popularized in an essay by philosopher Isaiah Berlin: "The fox knows many things, but the hedgehog knows one big thing." Those with a wide range of knowledge, and with a well-honed toolkit of approaches to solving these problems (the foxes), will invariably outperform those who have one idea and deploy it to solving all problems (the hedgehogs). Tetlock shows that those who have one grand theory that explains everything in the political world make poorer predictions than those who choose among many theories and find the one that applies best in a certain situation.

Gurung (2014, p. 111) also invokes the image of the fox and the hedgehog, in his case to discuss the role of methodology in the Scholarship of Teaching and Learning:

> *The fox devises many strategies; the hedgehog knows one great and effective strategy. If when trapped the hedgehogs' one strategy, rolling into a ball, does not work, it is all over. The fox with many alternative at her disposal fares better. Likewise, whether social scientist or humanist, the more strategies one has to End meaning, describe and explore learning, and examine the evidence, the better.*

Following from Gurung (and Tetlock), we can conclude that those who have more tools at their disposal can become better scholars of teaching and learning. The type of methodological trading Huber and Hutchings (2005) advocate, and that was practiced in my Carnegie residency and in places where scholars gather, serves our work very, very well, helping us to become more like the fox and less like the hedgehog.

For the Scholarship of Teaching and Learning to improve teaching capacity most broadly, we need to make the proverbial Big Tent truly welcoming to all kinds of methods. Every single research approach (experiments, close reading, student surveys, direct observation, etc.) brings with it certain flaws and biases. Most studies that use only one method can be criticized on methodological grounds; our conclusions are strengthened when we confirm our findings using multiple methods. The psychologist's surveys of student learning may not convince a humanist, while close reading of student essays may not convince the physicist. The critical issue should be alignment—the methods we use to assess learning must align with the kind of learning we hope to see. Studying scores on a multiple choice exam offers us evidence of one kind of learning, but it is hard to argue such scores reveal deep understanding of difficult concepts. In using multiple methods, like a fox, we must make sure that we choose the methods not because of our comfort level with them, but rather because they align with our teaching goals, allowing us to be more confident in our findings and convince a wider and more diverse audience.

Moreover, if we can use only one method, then we must be sure to be as transparent as possible in how we apply and explain it, and to be forthcoming about not only its strengths, but also its weaknesses. When I teach the drunk-driving article in my classes, I remind my students that whatever the article's external validity shortcomings, I have learned a tremendous amount from its internal validity strengths. The design is innovative, and the study provides the strongest corroboration of peer influence that I know of.[6] We can take what we have learned from this study, tuck it away in our heads, and move on to studies that build from what we have learned and address the weaknesses. Ultimately, as argued above, we are asking two questions: does the treatment (behavior of the confederate) cause an effect, and how generalizable are these results? Let us learn from studies that contribute to our knowledge about either of these questions, no matter what methods they used, and then build upon these results to design further studies that address the remaining holes in our knowledge. The perfect cannot be allowed to be the enemy of the good—if we only learn from any studies that are methodologically perfect, we will be waiting to become enlightened for a long time.

Within this context, we can address one persistent question within the Scholarship of Teaching and Learning: does this work require a control group? Grauerholz and Main (2013) are critical of the use of control groups in SoTL research because of the challenges found creating legitimate control groups, due to the nature of the classroom as a research setting. Every class, they argue, has its own personality and tone. They highlight this with Grauerholz's experience teaching "Women in Contemporary Society," a course in which she was trying to introduce student-centered approaches in one section but not in another. The classes differed in their general dynamic, quite apart from the presence or absence of these student-centered approaches, thus rendering comparisons tenuous at best.

How do we address this critique? Returning to internal validity, random assignment of students into control and experimental groups can be a valuable way of achieving internal validity. If, as in the drunk-driving study, we are able to control for all extraneous factors, isolating only the one factor of interest, our ability to say A caused B (and thus that we have internal validity) increases. Since we want internal validity, this type of research design should appeal to us, and we should look with favor upon studies that are able to leverage experimental and control groups to help achieve confidence that the treatment is causing the effect.

However, as Grauerholz and Main demonstrate, simply teaching the same class twice, with two different methods, does not automatically make for an effective experimental design. The classes may be different along other dimensions; even a subtle difference between teaching one class at 11 a.m. and the other at 2 p.m. may render the comparison ineffective, as might the presence of one particular student (for good or not-so-good) in one class and not the other. This is to say nothing of "control groups" taught by a different instructor, which are inherently problematic for claiming different results based on different course designs (since the difference may come from the instructor as well as the pedagogical method). Further, there may well be ethical issues associated with this approach: if I believe I have a superior method, is it ethical to teach one group of students using the seemingly inferior method in order to satisfy the methodological gods within the SoTL world?[7] The control group, often cited as a vital piece of SoTL methodology (Maurer, 2011), ought not always be treated as such.

When done well, a research design with a control group can be a compelling way to determine the effectiveness of a particular teaching innovation. Where such designs can be implemented ethically and effectively, we would be fools not to use them.[8] But given all the challenges with this, we would also be too rigid to demand that control groups be present for all good SoTL. The experimental design ought to be a tool that scholars of teaching and learning keep in their tool belts. Just like a hammer may be effective for solving certain problems but completely ineffective for others, the experiment has its uses, for which it ought to be valued. Nevertheless, if the goal is to demonstrate the effects of a particular technique, we ought not to be so rigid as to fail to see, and value, other potential methods for doing so.

EXTERNAL VALIDITY: GENERALIZATION AND THE PRESERVATION OF CONTEXT

Grauerholz and Main (2013; see also Booth & Hyland, 2000) are correct in pointing out that a concern with generalizability underlies a great deal of research in the Scholarship of Teaching and Learning. In describing the concept of generalizability as "inappropriate" (p. 157) to the SoTL literature, however, I believe they push the challenges with achieving generalizability (about which they are correct) too far in the direction of minimizing its importance. That generalizable results are difficult to obtain does not mean they are not desirable, nor does it mean that achievement of such goals should not be a standard by which we measure work in the Scholarship of Teaching and Learning.

Any time we begin to discuss generalizability, the notion of comparing apples and oranges arises. So, if Mark succeeds with a new discussion-generating technique in a small, upper-level poetry class at a highly selective university in the Midwest, would Cheryl be able to use this technique in her large introductory sociology class at a community college in California? I do not know. We might be tempted to say these situations are completely different, and that Mark's innovations could not speak to Cheryl at all. However, students are students, classes are classes, and it is at least theoretically possible that what

Mark has learned *could* be useful to Cheryl. Apples and oranges are not completely different—both are healthy food choices that could feed you if were hungry, for example.[9] While not all SoTL studies generalize outside the narrow settings in which they are performed, we ought to evaluate teaching innovations ("what works" questions) both in terms of how well they work, and in terms of how broadly we can apply their findings.

Generalizing happening effectively, ultimately, is a two-way street. As the author of a SoTL study, I should be expected to provide rich context (à la Shulman, 2013) of the educational setting in which I teach, the teaching strategies I use, and the learning my students have experienced. This attention to context shows us that no two classes are exactly the same, which makes generalizing from what I do in my class to what you can do in yours quite difficult. Yet, if we worship the unique aspects of our contexts too much, we lose the ability to learn from one another, and to ever allow our teaching to travel (Huber, 2009). To compound the generalizing problem, my ability to help others generalize is limited by the fact that I may not understand their unique context (their students, their discipline, or their course). I cannot help someone generalize to a setting I do not know intimately. All I can do is to provide enough information to make generalization possible; the job of pulling out the useful parts of my work and appropriating it for a different context rests at least as much upon the reader as the author. Such generalizations, when done carefully and intelligently, is how our field will grow.

CONCLUSION: DINNER WITH THE FAMILY

Recent literature in the Scholarship of Teaching and Learning has addressed the challenges of methodology in the field. While we have broad agreement on the necessity for appropriate and sound methodology in backing up our claims about student learning, disagreement continues as to what constitutes appropriate methods. Most commonly, we see some claiming that more humanities-driven methods are not scientific enough, and others claiming that the "rigor" demanded by those in the social sciences are inappropriate for the kinds of questions we are asking.[10] These methodological disputes are common in all fields, and are most likely a healthy development in a new and relatively unbounded movement within higher education.

What is troubling about many of these methodological debates is how, in the end, we are largely advocating for the same thing. The language of internal and external validity is intended to suggest that, ultimately, our work in the Scholarship of Teaching and Learning rests on these two central questions: Is what I am doing having an impact? Moreover, what can others learn from the work I am doing? If we are unable to determine if what we are doing is working, we exist in an evidence-free zone in which we are grasping in the dark to find the most effective ways to teach our content. In addition, if we are unable to generalize our work to other contexts, we are not building a field, and are not allowing the practice of teaching to advance outside our individual classrooms. Improving teaching across the university, as we know, is vital to the SoTL movement. Our movement suffers, immeasurably, if what any individual does in their own classroom is closed off to others in the field because nobody will pursue generalization.

In this essay, I have used the language of a family dinner to denote a group that may be united by common interests and goals, and yet be separated by differences simmering below the surface. I propose the language of internal and external validity to argue that, at the most basic level, we are all united by common questions. And, like a family, I propose this language to suggest that we might consider these common interests and cut each other a little more slack.[11] (Perhaps I am asking more of my SoTL family than most real families can manage!) Let us judge the work we do by how well it helps advance student learning (the internal validity side) and by how much others outside the narrow context can learn from it (the external validity side). Within that considerable space which unites us, this essay is a plea to be more open to different ways of reaching our shared ends.

ACKNOWLEDGMENTS

Much earlier versions of this work were presented at the 2013 ISSOTL Conference (Raleigh, North Carolina), and at the 2016 SOTL Academy Conference (Findlay, Ohio). I am grateful to Sarah Ginsberg, Matt Kaplan, Andrea Kaston-Tange, Nancy Chick, and especially Matt Fisher, who read this manuscript and offered valuable feedback. Pat Hutchings and Lee Shulman have provided guidance on methodology issues in the Scholarship of Teaching and Learning, which have significantly influenced my thoughts, and I am grateful for their wisdom. My work reflects countless conversations about research methods with friends and colleagues, too numerous to name, and insights I gained from them. All errors of fact and interpretation remain my own.

Jeffrey L. Bernstein is Professor of Political Science and Affiliated Faculty in Jewish Studies at Eastern Michigan University (USA). He is co-editor of Citizenship Across the Curriculum *(Indiana University Press, 2010) and of* Enhancing Teaching and Learning through Collaborative Structures *(Jossey-Bass, 2017).*

NOTES

1. Given that methodological disputes often define what is, and is not, “acceptable” work within a field, these differences often are part of the contested space within a discipline.
2. The classic tension between qualitative and quantitative research lurks within these discussions, but I would be hesitant to reduce these differences solely to that dimension.
3. Chick (2014) reframes visions of the possible as “What’s possible?” to make the construction of the questions similar, and to turn this idea into an actual question.
4. For example, my first published SoTL project (Bernstein & Meizlish, 2003) was done before I even knew what SoTL was—I was instinctively drawn to gathering data as a means of evaluating whether a particular teaching innovation “worked,” long before I realized there was a community of scholars engaged in such practice. I was initially motivated solely by seeing if my innovative teaching approach was successful.
5. Tetlock’s work explores the accuracy of predictions made by experts across a range of political issues. My focus here is on effective teaching practice, and the degree to which one can develop expertise in assessing it. While the contexts are different, the larger point about the value of being a fox versus a hedgehog applies across both.
6. As a technical point, this study does not demonstrate *peer pressure,* since even subjects with no confederates in their treatment entered the car ten out of ten times. Instead, it demonstrates the positive effects of *modeling,* since the only people who did not get into the car were those who followed the confederate in declining.
7. This is loosely analogous to the ethical question of withholding potential life-saving drugs from people in order to carry out double-blind control-group experiments. While such withholding of treatments is challenging ethically, it is justified by the concern that a bad drug can do serious damage to people if not thoroughly tested. Such fears are generally not present in the teaching and learning world, thus suggesting that using control groups may be more problematic in our work than it is in medicine.
8. For example, if an instructor has no prior belief that Method A would be better or worse than Method B, she could perform an experiment with no ethical qualms. Additionally, while a perfect, double-blind control study would be almost impossible to implement, experiments that fall slightly short of true internal validity would still help us gain a deep understanding of student learning; the quality of a control-group study should be conceived along a continuum, rather than as an

either/or. Shulman’s (2013) admonition to provide a rich, descriptive context in sharing our results would allow the researcher to see how close the research design comes to being a true experiment. Therefore, if Instructor A teaches the experimental group at 9 a.m. and the control group at 10 a.m., this would be close to a true experiment, subject to the caveats above. However, if Instructor A teaches the experimental group in the morning, while Instructor B teaches the control group at night, this would be too far from using a true control group to reasonably be called an experiment.

9. In teaching my classes, I use the phrase apples and oranges” versus “apples and tractors” to demonstrate that while some comparisons stretch too far, many times cases we think are quite different from one another turn out to have a far bit in common.
10. At the very least, these claims call attention to the fact that different scholars of teaching and learning likely have different meanings in mind when using a loaded term like “rigor.”
11. Chick’s (2014, p. 10) suggestion that we “be gentle with each other” strikes me as an appropriate way in which to approach these methodological controversies and disagreements.

REFERENCES

Bass, R. & Linkon, S. L. (2008). On the evidence of theory: Close reading as a disciplinary model for writing about teaching and learning. *Arts and Humanities in Higher Education, 7*(3), 245-261. https://doi.org/10.1177/1474022208094410

Bennett, C., & Dewar, J. (2013). SoTL and interdisciplinary encounters in the study of students’ understanding of mathematical proof. In K. McKinney (Ed.), The *Scholarship of Teaching and Learning in and Across the Disciplines* (54-73). Bloomington: Indiana University Press.

Bernstein, J. L. & Meizlish, D. S. (2003). Becoming Congress: A longitudinal study of the civic engagement implications of a classroom simulation. *Simulation & Gaming, 34*(2), 198-219. http://journals.sagepub.com/doi/10.1177/1046878103034002003

Booth, A., & Hyland, P., (Eds). (2000). *The Practice of University History Teaching.* Manchester: Manchester University Press.

Boyer, E. L. (1990). *Scholarship Reconsidered: Priorities of the Professoriate.* Stanford: Carnegie Foundation for the Advancement of Teaching.

Burns, H. L. (2016). Learning sustainability leadership: An action research study of a graduate leadership course. *International Journal for the Scholarship of Teaching and Learning, 10*(2), Article 8. https://doi.org/10.20429/ijsot1.2016.100208

Chick, N. L. (2014). ‘Methodologically sound’ under the ‘big tent’: An ongoing conversation. *International Journal for the Scholarship of Teaching and Learning, 8*(2), Article 1. https://doi.org/10.20429/ijsotl.2014.080201

Chick, N. L. (2013). Difference, privilege, and power in the Scholarship of Teaching and Learning: The value of humanities SoTL. In K. McKinney (Ed.), *The Scholarship of Teaching and Learning in and Across the Disciplines* (15-33). Bloomington: Indiana University Press.

Chick, N. L. (2006). From community property to public property: Shulman challenges CASTL participants to fill in moats and lower drawbridges. *ISSOTL International Commons, 1*(1), 7. http://www.indiana.edu/~issotl/newsletter/International_Commons_I.pdf

Felten, P. (2013). Principles of good practice in SoTL. *Teaching and Learning Inquiry, 1*(1), 121-125. https://doi.org/10.20343/teachlearninqu.1.1.121

Gale, R. A. (2005). Aesthetic literacy and the ‘living of lyrical moments.” *Journal of Cognitive Affective Learning, 2*(1), 1-9.

Glassick, C. E., Huber, M. T., & Maeroff, G. I. (1997). *Scholarship Assessed: Evaluation of the Professoriate.* San Francisco: Jossey Bass and the Carnegie Foundation for the Advancement of Teaching.

Grauerholz, L. & Main, E. (2013). Fallacies of SoTL: Rethinking how we conduct our research. In K. McKinney (Ed.), *The Scholarship of Teaching and Learning in and Across the Disciplines* (152-168). Bloomington: Indiana University Press.

Gurung, R. A. R. (2014). Getting foxy: Different magisteria in the Scholarship of Teaching and Learning. *Teaching and Learning Inquiry, 2*(2), 109-114. https://doi.org/10.20343/teachlearninqu.2.2.109

Gurung, R. A. R., & Schwartz, B. M. (2009). *Optimizing Teaching and Learning: Practicing Pedagogical Research.* Malden: Wiley-Blackwell.

Huber, M. T. (2009). Teaching travels: Reflections on the social life of classroom inquiry and innovation. *International Journal for the Scholarship of Teaching and Learning, 3*(2), Article 2. https://doi.org/10.20429/ijsotl.2009.030202

Huber, M. T., & Hutchings, P. (2005). *The Advancement of Learning: Building the Teaching Commons.* San Francisco: Jossey-Bass.

Huber, M. T., & Morreale, S. P. (Eds). (2002). *Disciplinary Styles in the Scholarship of Teaching and Learning: Exploring Common Ground.* Washington, DC: American Association of Higher Education and The Carnegie Foundation for the Advancement of Teaching.

Hutchings, P. (2000). *Opening Lines: Approaches to the Scholarship of Teaching and Learning.* Palo Alto: Carnegie Foundation for the Advancement of Teaching.

Hutchings, P. & Shulman, L. S. (1999). The scholarship of teaching: New elaborations, new developments. *Change, 31*(5), 10-15. https://doi.org/10.1080/00091389909604218

Lewis, C., Perry, R., & Murata, A. (2006). How should research contribute to instructional improvement? *Educational Researcher, 35*(3), 3-14. https://doi.org/10.3102/0013189X035003003

Maurer, T. W. (2011). On publishing SoTL articles. *International Journal for the Scholarship of Teaching and Learning, 5*(1), 1-2. https://doi.org/10.20429/ijsotl.2011.050132

Nelson, C. E. (2000). Effective strategies for teaching evolution and other controversial subjects. In J. W. Skehan & C. E. Nelson (Eds.), *The Creation Controversy and the Science Classroom* (19-50). Arlington: National Science Teachers Association.

Powell, J. L., & Drucker, A. D. (1997). The role of peer conformity in the decision to ride with an intoxicated driver. *Journal of Alcohol and Drug Education, 43*(1), 1-7.

Salvatori, M. R., & Donahue, P. (2005). *The Elements (and Pleasures) of Difficulty.* New York: Pearson.

Shulman, L. S. (2013, October). Situated studies of teaching and learning: The new mainstream. Keynote address at the meeting of the International Society for the Scholarship of Teaching and Learning, Raleigh, NC.

Tetlock, P. E. (2006). *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton: Princeton University Press.

Wineburg, S. (2001). *Historical Thinking and Other Unnatural Acts: Charting the Future of Teaching the Past.* Philadelphia: Temple University Press.