Mock Jury Research: Where Do We Go from Here?

Richard L. Wiener*, †, Daniel A. Krauss‡ and Joel D. Lieberman§

This paper reviews the four types of validity that make up Cook and Campbell’s traditional approach for social science research in general and psychological research in particular: internal validity, statistical conclusion validity, external validity, and construct validity. The most important generalizability threat to the validity of jury research is not likely a selection main effect (i.e., the effect of relying solely on undergraduate mock jurors) but is more likely the interaction of sample with construct validity factors. Researchers who try to capture the trial process with experimental paradigms may find that undergraduate mock jurors react differently to those efforts than do more representative community samples. We illustrate these issues with the seven papers that make up this volume, and conclude by endorsing Diamond’s call for a two-stage research process in which findings with samples of convenience gradually add more realistic trial processes and representative samples to confirm the initial findings and increase the research program’s credibility. Copyright © 2011 John Wiley & Sons, Ltd.

It has been over 30 years since Thomas Cook and Donald Campbell (1979) published their seminal volume on research methodology. Arguably, this treatise, which outlines “dos” and “don’ts” for social science researchers working in applied settings, is still the authority that many look to for understanding the strengths and limitations of research in applied social, organizational, and educational psychology. Indeed, in their well-used and often-cited case book that examines the role of social science in litigation, Monahan and Walker (2009) still rely heavily on Cook and Campbell’s basic methodological discussion to orient law and psychology students to the field of applied research in legal settings. For some reason the valuable lessons contained within the pages of this important monograph, which continues to echo its message throughout the program evaluation and applied research literatures in areas outside legal psychology (Shadish, Cook, & Campbell, 2002), have only had a limited impact on the area of jury decision-making research. The importance of the philosophy of research that Campbell and Stanley first presented in 1963, upon which Cook and Campbell elaborated 16 years later, partitions methodological issues into four broad categories of errors that researchers can make, each of which may limit the interpretability of their work. We review them below.

* Correspondence to: Richard L. Wiener, Department of Psychology, 238 Burnett Hall, University of Nebraska at Lincoln, Lincoln NE 68588.
E-mail: rwiener2@unl.edu
†University of Nebraska at Lincoln, Lincoln, NE.
‡Claremont McKenna College, Claremont, CA.
§University of Nevada, Las Vegas, NV.

Contract/grant sponsor: National Science Foundation; contract/grant numbers: SES-0096347.

Copyright © 2011 John Wiley & Sons, Ltd.
INTERNAL AND STATISTICAL CONCLUSION VALIDITY

The pinnacle goal of research in the behavioral sciences is to draw causal inferences between social, situational, psychological, and biological constructs on one side of the equation, and behavioral or social constructs on the other side of the equation. Cook and Campbell (1979) refer to internal validity as the property or ability of a research design to rule out all plausible rival hypotheses (Cook & Campbell, 1979) and allow interpretation of the results so that the researcher can infer a causal connection between the independent and dependent variables.

To increase the internal validity of studies in the area of jury decision-making research, we use experiments and quasi-experiments that rely on random assignment, experimental control, within-subject comparisons and the like. Indeed, the elegance of the simulation methods that researchers have developed to study mock jurors and mock juries in the last 30 or so years is impressive. Furthermore, each of these research designs uses sophisticated statistical techniques to analyze the relationships between variables in jury research, including recent advances in multi-level and structural equation modeling to increase the ability of a research design to draw reliable and powerful statistical inferences to test our causal hypotheses. Making advances in this later area of research, which Campbell and colleagues refer to as statistical conclusion validity, is also an area in which jury researchers have excelled. It is the desire to increase both internal validity and statistical conclusion validity that has driven many researchers into the laboratory to utilize mock jury paradigms, which allow greater experimental and statistical control. Indeed, reviewers of jury research spend a great deal of effort evaluating the internal validity and statistical conclusion validity of research submitted to top-level journals for the purpose of publication. It is fortunate for jury researchers that the kinds of practices (e.g., random assignment, experimental control, large sample sizes, and standardized instructions) that increase the internal validity also strengthen statistical conclusion validity. That is, internal and statistical conclusion validity correlate positively across research designs.

Ultimately, internal and statistical conclusion validities are largely problems of deductive logic. In fact, Cook and Campbell applied deductive logic to a variety of familiar research designs, to come up with lists of threats to validity that experimenters must either control or rule out in order to strengthen the causal inferences in those designs. For example, jury researchers randomly assign participants to experimental conditions conducted simultaneously to control for selection, maturation, and historical confounds. For many of us, the puzzle of designing a creative experiment within the constraints of deductive logic, allowing for strong causal inference, contributes significantly to the excitement of conducting jury research.

If the goal of our science was to study more general rather than more specific processes, that is, if we were interested in how people make moral, legal, and evaluative decisions in the abstract, that might be the end of the discussion. However, jury researchers are interested in how people make decisions in a much more contextualized environment. We are interested in learning how jurors, constrained by rules of court procedure and rules of evidence, make decisions and judgments that are moral and evaluative in nature, but that are determined by the elements of law, burdens of proof, and standards of evidence. That is, we do not really want to know how college students enrolled in introductory psychology classes read fact patterns about criminal and civil complaints and make attributions of causality. To be sure, the
process by which people make attributions of causality and responsibility is part of what jurors do in court but they do it under very different conditions and with many fewer degrees of freedom than researchers afford undergraduate students in psychology experiments.

**MAIN EFFECTS AND INTERACTIONS OF CONSTRUCT AND EXTERNAL VALIDITY**

To address these issues in applied research, Cook and Campbell (1979) introduced the notion of construct validity, which is the close cousin to, but not the same as, external validity.

External validity, of course, is the ability of a research design to generalize beyond the specific samples, times, and places of the original study. External validity is the focus of this special edition because, at the most fundamental level, we were interested in whether mock jury data collected with college students as participants generalize from student samples to populations of jury-eligible citizens. Unlike internal and statistical conclusion validity, the issue of external validity is not really a problem of deductive logic but is instead an issue of inductive logic or, more familiarly in the language of research psychologists, an empirical issue. In the end, the only way to determine whether results obtained with a student sample generalize to other samples (e.g., eligible jurors in the community) is to compare the outcomes and processes of the decisions that students make with those that eligible jurors would make when evaluating a criminal or civil fact pattern.

However, as implied above, at least as important and perhaps even more important is the fourth standard, construct validity, which Cook and Campbell (1979) introduced to measure the quality of research designs in applied settings. Construct validity is concerned with the measures, manipulations, control groups, and the authenticity of the research designs in relationship to the constructs that the researchers intended to sample, measure, manipulate, and, in the case of mock jury research, simulate in the real world. In other words, construct validity asks whether the research was successful enough in translating social and psychological reality into the research paradigm to be able to interpret the data to make statements about the social and psychological reality that generated the research hypothesis in the first place. This type of validity is as important, and perhaps even more important, than other forms of validity.

The relationship between external and construct validity is at the heart of the debate about whether or not college students are an adequate stand-in for jury-eligible citizens. Construct validity is in fact a specific type of external validity. It is concerned with the extent to which the research measures and manipulations generalize beyond their specifics to the social and psychological conditions they purport to represent. In jury research, the question that we need to ask, and too often fail to ask, is whether our laboratory experiments adequately sample the essential elements of the law to allow us to make statements about how jurors are likely to make decisions of guilt, liability, and damages. For example, researchers must be concerned with whether the internal validity and statistical conclusion validity design features, such as pretest measures, multiple verdict assessments, truncated fact patterns, and the lack of deliberation, do serious damage to the construct validity of research designs. For these reasons, some researchers (Vidmar, 2008) are skeptical about the contributions that mock jury research can make.
to the science of jury decision-making. Some would argue that once an experimentalist
disassembles the gestalt of the courtroom in the laboratory for purposes of analysis, the
researcher will never again be able to generalize back to the courtroom. This, too, is an
inductive question, which we cannot answer with logic alone, but instead it requires an
empirical approach that compares results from simulated mock jury research to actual
trials. In fact, Cook and Campbell (1979) and Shadish et al. (2002) make the convincing
case that construct and external validity correlate in a positive direction. Most
importantly for mock jury research, it would seem that the very mechanisms in research
that increase generalizability (i.e., external and construct validity), such as studying
natural samples of jurors who evaluate real cases in real or nearly real courtroom
contexts, tend to lower the internal and statistical conclusion validity of research designs.
In short, internal and statistical conclusion validities correlate inversely with external and
construct validities. As a result, researchers sometimes find themselves in the
unfortunate position of buying increased power in causal inference about how jurors
reach judgments in civil and criminal cases at the cost of the ability to generalize the
findings to real jurors making real decisions in real trials.

Our argument picks up here and goes one step further. It seems to us that in jury
research, the issues of external validity and construct validity might be even more highly
interconnected than research methodologists imagine. That is, when researchers ask
jurors to make civil liability decisions or criminal culpability decisions, the decision
changes not only as a function of the jury instructions, quality of the jury simulation,
elements of the law, admissibility of evidence, and the like, but also as a function of the
nature of the decision-maker. That is, the issue for us is about the interaction between the
measures and manipulations and the type of individual reaching those decisions. We are
hardly the first jury researchers to consider the interaction of people with measures and
manipulations. At the same time that Cook and Campbell were writing their seminal
paper and elaborating on the earlier Campbell and Stanley approach to different types of
validities, Weiten and Diamond (1979) were applying these same ideas to jury research.
Weiten and Diamond surveyed the then existing simulated jury research and identified
six threats to external validity and policy relevance. These included (not in their original
order) inadequate trial simulations, lack of jury deliberations, inappropriate dependent
variables (focusing too often on criminal punishment and lacking actual elements of
charged crimes), lack of corroboration with field research (i.e., convergence of findings
between laboratory and field studies), problems in role playing (i.e., the difference in
consequences of decisions for real and mock jurors are substantial), and inadequate
sampling (relying too heavily on college student mock jurors).

**EMPIRICISM CONSIDERATIONS AND LIMITS TO THE MOCK JURY PARADIGM**

The first five of these categories are problems in construct validity but the last is clearly an
issue of what we more traditionally refer to as external validity. Furthermore, when
discussing the issues inherent in inadequate sampling issues, Weiten and Diamond
(1979) consistently entertained the likelihood that “interactions between experimental
treatments and subject variables may pose threats to valid cross-sample inference.”
(p. 76). For example, Weiten and Diamond (1979) considered an early study in which
Berg and Vidmar (1975) showed that conservatism correlated positively with harsh jury
decision-making and argued that student mock jurors may be less conservative, less authoritarian and therefore more likely to be lenient in criminal trials. Another way of thinking about this problem is that juror attitudes interact with case facts, charges, and trial procedure in such a way as to threaten the construct validity of the studies that rely solely on undergraduate mock jurors. Put bluntly, the question is simply whether undergraduates possess attitudes and beliefs that alter their perceptions of legal constructs, such as the meaning of the elements of crime or a tort, beyond a reasonable doubt versus preponderance of the evidence, criminal intent, negligence, burden of proof, and the like, relative to those of jury-eligible community residents. If so, then the interaction of type of participant and type of procedure may indeed become a subtle but important threat to the generalizability of mock jury research. Weiten and Diamond (1979) go on to suggest that other factors, such as differences in cognitive capacities, might make student jurors different from jury-eligible citizens. If students have greater memory capacities and information-processing abilities than do typical jurors, then they may reach very different judgments about the same case facts, or even if they reach similar judgments, they may do so using very different approaches, styles, and information. Indeed, in an independent assessment of the simulated jury research, Bray and Kerr (1979) shared Diamond’s concerns (Weiten and Diamond, 1979) and called for researchers to systematically explore how well simulated research studies generalize to real trial results. Some, but not all, argue that this work is still lacking in the literature.

Eighteen years later, Diamond (1997) revisited her arguments about the threats that mock jury research present to external validity and policy relevancy and concluded that some progress was evident in the literature. Her review showed more studies with realistic measures, legally relevant trial simulations, and even more use of jury-eligible mock jurors. Still she found room for improvement, grounded again in concerns about the interactions between participant attributes and legal processes. She opined that there were mixed results in the literature comparing student samples with eligible jurors, with some showing parallel findings (Casper, Benedict, & Perry, 1989) but others finding major differences between the samples (Severance, Greene, & Loftus, 1984). The problem in 1997 remained the interaction between external and construct validity. According to Diamond (1997), we had no theories to help us understand when and how differences between student and general adult populations interact with the experimental manipulations that researchers used to try to capture the jury experience in a simulated mock jury study. Diamond (1997) called for efforts at building a database to compare the differences between student samples and adult samples and to develop theories that guided researcher discretion about sampling. Diamond concludes that, “…until we can predict reliably when a student sample is likely to provide an adequate model of juror behavior, jury venires should remain the preferred source for maximizing both face and external validity.” (p. 563).

Shortly after Diamond revisited this issue, Bornstein (1999) also addressed the concerns of external and construct validity in mock jury research. Bornstein took seriously the need for an empirical assessment of the relationship between students’ and jury-eligible citizens’ performance in mock jury trials. He collected and displayed a table of 26 jury decision-making experiments, which included student and non-student samples, and tabulated how many studies found main effects that distinguished student and non-student samples and how many found interactions. While this analysis does not constitute a true meta-analysis with an exhaustive list of studies, measures of effect sizes and fail-safe statistics, Bornstein found only six out of the 26
that reported main effects for type of participant and fewer that found interactions with other study factors. He suggests that jury researchers can take solace from the findings that type of participant does not seem to matter a great deal in mock jury research. Indeed, many articles published after 1999 cite this paper to justify the use of convenient but less than realistic simulations and undergraduate participants in mock jury research.

While we welcome the Bornstein analysis as a first step in examining systematically what we know about student mock jurors and community samples, it does not address several important factors, including suitable attention toward the specific types of interactions among construct validity issues in jury research (burdens of proof, standards of evidence, deliberations, types of charges, and the like). For example, at even the most elementary level, we wonder whether type of charge matters and we suspect that it does. Very little, if any, data directly compare differences in how student and community jurors think about facts and evidence even within criminal cases with a common theme, such as violent crime. Similarly, this type of important validation research is largely absent from studies of civil jurors and juries.

Despite researchers’ hopes that juror type has minimal consequence, there are, as of yet, no formal meta-analyses that we know of that tested differences in samples or interactions between samples and other study factors in mock jury research, with the exception of Mitchell, Haw, Pfeifer, and Meissner (2005), who examined race effects in mock jury research. The authors studied both jury verdicts and sentencing and found small effects for racial bias more among Black than among White research participants. For our purposes, the most interesting finding from this study was that the type of participant moderated the effect of the race of the defendant, so that community members were more likely to show racial bias in sentencing (but not for verdict decisions). Thus for at least some purposes there is some formal evidence that type of juror does modify judgments about defendants. Where else will we find evidence of interactions between external and construct validity factors?

The findings of the Mitchell et al. (2005) study are certainly important but perhaps equally important is the difficulty that the researchers had in studying the moderating effects of type of sample. For the verdict decision, the authors located 37 experiments that studied the effects of race in student samples, but only seven that did so with community samples. For the sentencing decision, there were 15 student sample studies, but only five community sample studies. In short, the Mitchell et al. (2005) meta-analysis suggests that the interaction between defendant race and type of sample is worth further investigation but that until the literature includes more studies that compare community samples with student samples, formal comparisons using meta-analytic techniques to test this and other interactions between external and construct validity threats will remain difficult.

**FACTORS TO INCREASE REPRESENTATIVE RESEARCH**

With the advent of newer technologies, it will be easier for researchers to extend their samples to include more participants that represent jury-eligible participants. It may not be as inconvenient or as inefficient as it once was to recruit representative samples of mock jury participants. More and more courts and jury pool administrators are willingly opening their doors to psychological researchers. In addition, the internet has
opened up a variety of new means to gather more representative samples using tools such as Craigslist to solicit participants to come to a research lab. Alternatively, researchers may take advantage of online subject pools such as Amazon’s Mechanical Turk (www.MTurk.com), which allows researchers to post simulations and surveys to a significantly more diverse internet participant sample than college students for limited cost and data collection time.

One recent study of Mechanical Turk found acceptable psychometric standards for a host of personality measures as well as a significantly more diverse sample of participants than other internet derived participant samples (Buhrmester, Kwang, & Gosling, 2011). We are by no means suggesting that these procedures for participant sampling should replace traditional sampling techniques, or that these new procedures do not contain their own limitations. However, we are hopeful that technological advances in sampling will facilitate the collection of more representative samples in a more efficient and less expensive manner so that it will be easier to compare student samples with community-based samples.

With increases in the heterogeneity of sampling, it will become possible for journals to publish studies that compare findings with multiple samples or that replicate research where the external validity of the original study was limited by the use of college student mock jurors. Although psycholegal journals are reluctant to publish replications where the key difference between the initial publication and the replication is a sample difference, this is an approach that the health sciences have used for many years, starting with the publication of animal studies and culminating in the publication of random treatment trials with human participants. In addition to studies that compare samples, those that replicate findings with samples closer and closer to the population of ultimate interest are highly beneficial because each replication allows for greater confidence in the treatments used. We are hopeful that by adopting an approach that compares samples or publishes replication studies, we can better understand sample differences and convince the courts to place more confidence in psychological interventions that facilitate more effective jury decision-making.

SEVEN TREATISES ON CONSTRUCT AND EXTERNAL VALIDITY IN JURY RESEARCH

We have only begun to address these issues and there is much more to do. This special edition was the beginning of a systematic effort to probe the specificity and generalizability of the mock jury paradigm, which features student mock jurors at its center. The papers in this edition raise many of the questions but only begin to offer answers. We need to determine the conditions under which jury simulations with undergraduate mock jurors reliably model the types of decisions that real jurors make in a courtroom environment. The task is certainly daunting and that is perhaps why some researchers prefer to dismiss its importance.

We believe that there is a need to work at multiple levels to address this fundamental concern. First, the articulation of theory that explains how manipulations of trial processes, types of evidence, attorney behavior, juror attributes, motivational states, and emotional states impact judgments and decisions will allow us to understand the conditions under which generalization is likely to be successful or not successful. It is difficult to overstate the importance of strong theory to guide the empirical tests of the
interactions among elements of construct validity (i.e., jury process) and external validity (e.g., type of sample). Theory in jury decision-making research could predict the types of studies where sampling effects are likely to occur, and when the use of students may be problematic. This allows researchers to design studies with a stronger methodological foundation, because researchers can make informed a priori decisions regarding when college students are an appropriate sample and when they should avoid using them as research participants.

Second, however, is the fact that ours is an empirical science and the ultimate determination of when it is necessary to test student jury results against a sample of jury-eligible decision-makers must rest upon data. As Bornstein’s (1999) review showed, there are not enough comparisons between samples of convenience and samples drawn from a community population of eligible jurors to test the generalizability of the theories and relationships that researchers attempt to model in their experimental work. Each of the papers in this volume illustrates these issues in their own ways.

McCabe and Krauss (2011) tested a correction for affective and cognitive biases among jurors in a civil commitment hearing for sexually violent offenders in two studies. In the first study, they found that student jurors demonstrated a leniency effect when attorneys acknowledged their mock jurors’ emotional reactions to the case materials. However, in a second study conducted with a community sample, this intervention based in theory to reduce emotional influences in jury decision-making was not nearly so successful. Furthermore, a sample of representative jurors, as compared with students, was less sensitive to differences in expert testimony based upon actuarial as opposed to clinical opinions. Finally, measures of cognitive effort predicted the jury behavior of the undergraduates but not that of the community respondents. It would seem that future research should carefully look for interactions between sample type with emotional reactions, receptivity to type of expert testimony, and the effects of effortful processing. This is an excellent example of how external validity concerns could jeopardize the construct validity of a simulation.

Next, Fox, Wingrove, and Pfeifer (2011) (the second paper) examined the allocation of damages in a medical malpractice case with both undergraduate and jury panelists serving as mock jurors and found that the students assigned lower punitive damage awards. These differences resulted in part from the fact that the panelists were more influenced by evidence that supported compensatory damages than were the student jurors; that is, the former showed more “bleeding” across the types of damage awards. In fact, the jury panelists were more inclined to assign the entire punitive award to plaintiffs while the students were more favorable towards a split recovery. The importance of type of sample for understanding split recovery shows, once again, that the construct validity of a simulation (i.e., the influence of evidence on decisions of how to apportion blame and damages) can depend upon whether students or community members act as mock jurors.

Keller and Wiener (2011) (the third paper) switched focus to criminal jurors and examined the differences between students and jury-eligible community members in their evaluations of culpability for highly emotional charges brought against defendants whom the state accused of either homicide or sexual assault. They found that the interaction of jury attitudes and type of sample produced very different results for both types of cases and inasmuch, documented how sample type can limit the generalizability of juror attributes in predicting case outcomes. Specifically, for
community members, but much less so for student jurors, evaluators who showed higher bias against defendants (i.e., higher scores on the jury bias scale and on attitudes toward rape) were most likely to find the defendants guilty of these sexual assault charges. Going in the opposite direction, findings from judgments of the homicide vignettes showed that student mock jurors were more lenient toward defendants than were community mock jurors. To the extent to which attitudinal differences co-vary consistently with sample type, research that uses student samples in highly charged cases could run the risk of creating a decision task that produces results that diverge from those that are likely to come out of criminal courtrooms where empaneled jurors hear these types of cases.

The fourth paper reported on mock jurors’ decision-making in medical malpractice cases in which the plaintiffs sued physicians who performed surgeries that went badly. Reichert, Miller, Bornstein, and Shelton (2011) compared the judgments of students and community members acting as mock jurors, varying the reason for surgery (medically necessary vs. elective), the type of surgery (gastric bypass vs. nasal reconstruction), and measured individual difference factors, such as whether participants held negative perceptions of overweight patients who undergo elective surgeries. Although the authors found individual differences between students and community participants in relation to these attributes, those individual differences did not produce results that consistently favored one side over the other. Instead, student mock jurors were more likely to find doctors liable, and were more likely to assign responsibility to patients than were members of the community sample. At the same time, jury-eligible community members found overweight patients less responsible for their situation than normal weight patients, but students showed the opposite pattern. Reichert et al. suggested that jurors might sympathize with patients and show defensive attributions in their perceptions, which could account for the differences between students and jurors in their perceptions of the overweight. These attributions are a form of self-serving preservation in which jurors refrain from blaming patients because they feel a sense of vulnerability. Future work examining medical malpractice cases would do well to examine the role of differential sympathy among students and community sample mock jurors as a partial explanation for results.

The fifth paper turned our attention toward verdicts in a Title VII discrimination case examining how jurors react to a Latina complainant making allegations of hostile work environment sexual harassment. Schwartz and Hunt (2011) varied the ethnicity of the complainant (European American or Latin American) and presented a mock trial vignette to undergraduates and to community member participants. The authors developed a cultural relativist argument to explain the behavior of the Latina complainant who did not report her concerns immediately to her supervisors. Delayed reporting is a factor that is part of many sexual harassment complaints, and which could support judgments that the plaintiff welcomed the ostensible offensive conduct. In this case, the ethnic background of the complainant was the reason she did not raise her concerns immediately to her supervisors. Importantly, Schwartz and Hunt found that while the cultural relativist argument did not affect undergraduates’ judgments, it caused a backlash among community members, leading to lower levels of defendant liability. Had the authors used only a student sample, they would have found that the inclusion of a cultural relativist argument had little impact. Here, type of sample interacted with the type of evidence in a rather dramatic manner, demonstrating that the typical undergraduate student sample may be more accepting of diversity-based
arguments, while community members, who likely populate juries, may react very strongly in the opposite direction.

Nuñez, McCrea, and Culhane (2011) approach the issue of construct validity from a very different point of view and argue that the bigger threat in the jury literature is that absence of deliberations in most mock jury studies. In this, the sixth paper in this issue, Nuñez et al. (2011) maintain that the main effect of selection (relying on undergraduate mock jurors) is of less concern than is the simple main effect of a key trial process, specifically the absence of jury deliberations. These authors would favor jury decision-making studies that employed deliberating panels of undergraduates over those that fail to include deliberations at all. They review the existing literature and conclude that the field can only fully appreciate the importance of the type of sample when researchers are able to compare the differences in deliberations among panels of community members and panels of undergraduates. They point out that for this issue, the interaction of sample with presence versus absence of deliberation is ultimately an empirical concern that requires data for a final resolution.

Indeed, in some cases, student mock jurors may represent real jurors as well as do community members. The problem for our science is determining the limits to generalizability so that we can distinguish situations in which students will suffice and when they will not. For example, Hosch, Culhane, Tubb, and Granillo (2011) in our seventh paper presented a videotaped criminal trial re-enactment to samples of jury-eligible university students and volunteers from the venire in the same community. The substantive factors of interest were the language in which the defendant testified (English or Spanish) and the race of the defendant (Northern European, Latino, and African American). While the authors found demographic differences separated the community sample and student mock jurors, these differences did not translate into differences in guilty verdicts or interactions between demographic variables and the factors of interest on guilty verdicts. In fact, verdict patterns were similar for student mock jurors and community participants. However, the students’ sentence recommendations were significantly and substantially more punitive. Hosch et al. demonstrate the importance of a theory of jury decision-making that takes into account the demographic characteristics of the jurors in order to be able to predict and control the effects of sample characteristics on jury decisions for culpability, liability, punishment, and damages.

CONCLUSIONS

Cook and Campbell (1979) provided us with a convenient and powerful way to think about the ability of our research results to generalize beyond the immediate data that we collect. The papers that make up this special edition show why it is time for jury researchers to take seriously the issues of external validity, and especially the way it limits construct validity in jury studies. It is important to note that, although we were interested in the issue of sample effects, we did not use the existence of sample effects as a criterion for either soliciting or accepting manuscripts for this issue. Rather, the criterion we used was simply whether a manuscript contained a direct comparison between college students and a more representative sample or commented on the importance of those comparisons. If these studies had relied only upon a sample of undergraduates, several of the authors would have found very different effects in both
civil and criminal studies. Indeed, some of the authors would have entirely missed the fine distinctions in the way in which mock jurors evaluate these types of cases and would have been unable to uncover those subtleties.

The use of undergraduate mock jurors is efficient and in many cases very suitable for the initial tests of important issues, even issues with some practical significance. For example, Rose and Ogloff (2001) showed that researchers could make extensive use of undergraduate participants to test the comprehensibility of jury instructions, pointing out areas where they need improvement and areas in which people understand the plain meaning that instructions try to convey. However, given the kinds of findings that we show here, it is still necessary to verify the final results of such testing with jury-eligible respondents. Wiener et al. (2004) documented high levels of miscomprehension in a community sample of mock jurors, which may have been lower in a sample of undergraduate college students who were more familiar with mastering new and complex information and taking tests on their understanding. Of course, specific examples like these highlight the importance of obtaining empirical answers to empirical questions. Our main point is that when there is some expectation that behavior will not be identical from sample to sample, jury researchers can only be sure of the generalizability of their findings when they compare results from different samples.

Vidmar (2008) commented that the most important structures to study in jury research are the effects of the judge’s preliminary comments, the attorneys’ opening statements, the direct and cross-examinations of plaintiffs’ and defendants’ attorneys, the actual testimony of the witnesses, the testimony of experts, attorneys’ closing arguments, and finally jury instructions. Each of these important structural issues could influence the substantive judgments that are the products of both civil and criminal trials. Vidmar complains, justifiably, that jury simulation studies often fail to examine the relationship between the structural elements of trials and the judgment outcomes. We argue that relying only on the results of studies that use undergraduate students as mock jurors compounds this problem because students and eligible jurors may view charges, trial process, and trial procedure from very different perspectives. As Diamond (1997) said, “…until we can predict reliably when a student sample is likely to provide an adequate model of juror behavior, jury venires should remain the preferred source for maximizing both face and external validity” (p. 563).

Along similar lines of analysis, Judge William Caprathe, who commented on the seven papers that make up this special edition, came to a similar conclusion (Caprathe, 2011), but he arrived there by way of a very different path. Judge Caprathe, a long-time trial attorney and current trial judge describes the cultural and educational value of the jury system and comments on recent efforts of the American Bar Association to promote some badly needed reforms. He argues that the jury enterprise as we practice it in the United States makes a unique contribution to the education of our citizens and to their opportunity to participate in their government. Therefore, attempts to improve the jury system are of the utmost importance to the political experience of our democracy. As such, these efforts must be taken seriously and therefore need to include the best, most unbiased research efforts available. Judge Caprathe recognizes the value of psychological research on juries and respects the contribution that these studies can make to both ongoing litigation and, perhaps more importantly, to the public policy debate. He, like several of the social science commentators that we referenced above, favors a two-stage process in which initial findings using mock jurors in minimalist paradigms are tested against the responses of community samples in
more ecologically valid simulation paradigms that include deliberations and realistic trial procedures. He states:

"To determine if changes in the jury system are needed and would be beneficial, mock juries using the innovations need to be compared to mock juries that are not. In addition, studies such as the ones conducted or reviewed by the authors of these seven articles are needed to determine the significance of the make-up of mock juries. Then, with that significance determined, studies testing the innovations will carry greater credibility.

All of the authors to some degree acknowledged that using students exclusively for mock juries causes differences in applicability to actual jury trials. The significance or even the extent of the significance differs depending on many other variables; for example, the type of case and the method used by the researcher. All agree that more research is needed before any absolute conclusion can be reached. A good solution in the meantime is the two-stage method so that if any of the results of the student mock jury are not applicable, then specific adjustments to the method of research and the conclusions can be made."

We conclude that, at the very least, jury researchers should apply different theoretical models to study multiple sample testing for convergence of findings across populations before they place too much confidence in findings with only undergraduate mock jurors. Furthermore, we advocate conducting studies that compare samples and studies that replicate findings with samples closer and closer to the population of ultimate interest. We are hopeful that by adopting an approach that compares samples or replicates early findings with limited generalizability, we can better understand sample differences and convince the courts to place more confidence in psychological interventions that facilitate more effective jury decision-making.

REFERENCES


