Opening the Black Box:
The Process for Selecting Non-experimental Evaluation Methods and the Impact on Postsecondary Education Programs and Policy

Elizabeth Vaade
Wisconsin Center for the Advancement of Postsecondary Education (WISCAPE)
University of Wisconsin–Madison
everade@education.wisc.edu

Bo McCready
Wisconsin Center for the Advancement of Postsecondary Education (WISCAPE)
University of Wisconsin–Madison
mccready@wisc.edu

Scholarly paper prepared for the Council on Public Policy and Higher Education (CPPHE) Forum at the Association for the Study of Higher Education (ASHE) Annual Conference
November 16, 2011

Abstract
This study examines the factors that influence the selection of non-experimental quantitative methods in postsecondary educational evaluation research. To explore the method selection process we reviewed existing research and engaged in conversations with experienced policy actors to better understand their perceptions of these methods. We identify the factors that help researchers determine which methods to use and how researchers and practitioners assess and interpret these techniques. We group our findings into four themes: (a) the value and volume of non-experimental research, (b) the factors that impact selection and assessment of methods, (c) the perceptions of specific methods, and (d) the tension between rigor and accessibility. Based on these findings, we generate points for future discussion that may help point the postsecondary educational research community toward larger questions about evaluation methods, perceptions, and value.

DRAFT – PLEASE DO NOT CITE WITHOUT AUTHORS’ PERMISSION
Introduction

As budgets get tighter and postsecondary institutions face intensifying challenges in serving students and key stakeholders, administrators and policymakers confront difficult choices about what programs to preserve, expand, or cut. The push from many directions for data-driven decision-making and accountability should place evaluation at the center of this process. Effective program evaluation relies on the development of an appropriate basis of comparison to help determine if programs are meeting their goals. These comparisons allow evaluators to better estimate program impacts through systematic analysis of an approximate *counterfactual* (i.e., what would have happened to participants in the absence of an intervention).

Because of ethical and practical concerns, it often is infeasible for educational administrators to randomly select participants and create a true experimental control group. To deal with these concerns, researchers and practitioners have explored non-experimental evaluation designs, including the creation of comparison groups, which approximate randomization by comparing program participants to substantially similar non-participants.\(^1\) There exists a great deal of research on non-experimental methods. However, much of this literature focuses on the results of one method and operates under the assumption that the reader understands that method. These articles pay little attention to the process researchers use to select a method, the tradeoffs inherent in that decision, and how policymakers and practitioners can interpret results derived from a particular technique. Without this information, rigorous evaluation can seem like a black box where inputs and outputs are visible but the processes remain hidden and largely inaccessible.

In this paper, we look to expand the discussion of evaluation research and method selection among postsecondary educational researchers and other policy actors by examining the factors that

---

\(^1\) In this paper, we use the term “non-experimental” to refer to any quantitative evaluation designs that do not involve randomized controlled trials or experiments. For our purposes, “non-experimental” also encapsulates designs that are sometimes referred to as “quasi-experimental.”
influence the selection of non-experimental designs. To do so, we reviewed existing research and engaged in conversations with a purposefully sampled group of experienced policy actors to better understand their perceptions of these methods. We identify the factors that help determine which methods to use and how researchers and practitioners differ in their interpretations of these techniques.

In fields such as college access and success, teaching and learning, and other student-centered initiatives, evaluation can play a critical role in helping educational administrators to add, improve, or eliminate programs and guide institutional allocation of resources. More appropriate methods provide better estimates of program effects. To increase effectiveness, researchers and other policy actors must understand each other’s process of selection and perceptions of these methods. By conducting this small-scale study, we hope to shed some light on these issues.

Theoretical Framework

In the past 10 years, randomized controlled trials have emerged as the “gold standard” in education research (Whitehurst, 2003), but they are not impervious to criticism. In theory, random assignment of treatment removes the potential problem of correlation between unobservable characteristics, program participation, and outcomes. As a result, it is much easier for researchers to make causal claims about program effects when using random assignment. The benefits of random assignment were first articulated by R.A. Fisher in 1928, and it soon became the norm for agricultural, biological, and medical evaluation research (Imbens & Angrist, 1994; Burtless, 1995).

The popularity of random assignment in education has increased as researchers, policymakers, and organizations push for more “scientific” research in education (Shavelson & Towne, 2002; Walters, 2009). However, randomized controlled trials are often infeasible for several reasons. Critics argue that random assignment often lacks external validity, poses serious ethical concerns, presents logistical
challenges in educational settings, and does not always capture the intent of a program accurately (Astin, 1991; Hellman & Hellman, 1991; Cook, 1999; Winship & Morgan, 1999). These concerns are compelling enough that many programs may be unwilling to practice random assignment. Therefore, evaluators often must adopt non-experimental designs for estimating program effects.

One common non-experimental approach involves constructing comparison groups of individuals who do not participate. Often, evaluators construct these groups using matching techniques to ensure that the comparison group is relatively similar to the participant group based on a set of observed characteristics. However, there is no universally agreed-upon best method for comparison group construction, and scholars use a variety of methods to create these groups, including using all non-participants within a population, creating pairs of participants and non-participants that are matched based on one or more variables or their propensity to participate in the program, and creating groups clustered around a cutoff point that determines participation.

Researchers who use non-experimental methods face their own set of problems. Because governmental funding and policy have favored some methods over others and contributed to the increased push towards “scientific” research, some educational researchers feel as though they live in a world where policy influences research, rather than the other way around (Lather, 2004). For example, the Institute of Education Sciences states that “Only well-designed and well-implemented randomized controlled trials (RCTs) are considered strong evidence (emphasis ours),” and that it is impossible for any non-randomized study other than those using a specific form of regression discontinuity or single-case study to meet their evidence standards “without reservations” (U.S. Department of Education, 2008; Kratochwill et al., 2010). As a result, a large body of research is immediately disregarded by an organization that bills itself as “the nation’s engine for education research, evaluation, assessment, development, and statistics,” creating structural and cultural barriers for researchers using certain methods.
Further, researchers who study non-experimental methods have not always provided concrete guidance on method selection for others performing and consuming evaluations. Some scholars have compared random selection to other methods of comparison group construction (Dehejia & Wahba, 2002; Michalopoulos, Bloom, & Hill, 2004), but very few, if any, have compared different non-experimental methods to each other. Stuart & Rubin (2007) recommend experimenting with multiple methods and selecting the one that creates the most appropriate comparison group. Although this advice is sound, it also requires knowledge and understanding of a variety of methods, along with the time and technical expertise to test them. Many researchers may possess this capability, but other policy actors may not have the necessary levels of expertise, particularly those who use the findings of this work to make policy and programmatic decisions.

The disconnect between research and policy plays into these questions of method selection and interpretation. The research-policy divide has been highly problematic for postsecondary education, and although much attention is paid to this problem, it appears to persist. In the educational research community, scholars recognize that research may not always influence policy design and implementation (Birnbaum, 2000; Ness, 2010). Scholars also have expressed uncertainty as to the role that information derived from policy analysis and evaluation plays in the policymaking process (James & Jorgensen, 2009). For instance, Gray & Lowery (2000) found that legislators and staffers in Minnesota made policy decisions based largely on their experiences and the experiences of their constituents, as opposed to research and analysis.

Because this disconnect appears to exist, it is not a stretch to suggest that researchers’ and policymakers’ understanding of certain evaluation methods might be very different, which serves as yet another barrier to research findings informing policy decisions and the two-way exchange of ideas between these groups. Similarly, if policymakers remain in the dark about the method selection process, they may be less likely to value evaluation research and more likely to misinterpret its findings. Finally, it
is worth exploring whether the disconnect comes from a lack of knowledge about non-experimental methods among policymakers or from a belief that this kind of work is not valuable when making policy decisions.

Motivation and Goals for Study

This project was undertaken by members of a research team from the Wisconsin Center for the Advancement of Postsecondary Education (WISCAPE), a postsecondary education policy research center located within the School of Education at the University of Wisconsin–Madison. WISCAPE acts at the intersection of the research and policy spheres to help policymakers, researchers, and other policy actors acquire and share knowledge throughout the postsecondary education policymaking process. To do so, WISCAPE positions itself as a translator, focused on making academic research more accessible to policymakers and practitioners and helping scholars better understand the realities of policymaking and practice.

In our time working at WISCAPE, we have served as both producers and consumers of evaluation research that uses non-experimental quantitative methods. We also have worked with a variety of policy actors—including postsecondary educational administrators, faculty, staff, legislators, government agency staff, and practitioners—who create and use this work regularly. In these experiences, we realized that these groups, at times, lacked understanding of the available options for evaluation and the method selection process. This motivated us to pose the following key questions that we examine in this study:

1. How valuable is non-experimental evaluation research to the creation of good postsecondary educational policy?
2. How much of this research currently exists? How much should exist?
3. How do producers of this research select their methods? How do consumers assess whether the methods chosen are appropriate?
4. How do various policy actors perceive specific non-experimental methods?
5. How do policy actors prioritize rigor versus accessibility? Are there tradeoffs between these goals?

In asking these questions, we hope to illuminate the method selection processes used by researchers as well as explore how policymakers and practitioners view and use non-experimental research. If the opinions of these policy actors do not align with one another, we will be able to provide potential reasons why non-experimental evaluation research is not reaching its intended audience or having the desired impact.

**Method of Inquiry**

Our goal was to obtain a qualitative perspective of these quantitative methods and elicit perceptions of non-experimental evaluation, specific methods, and other areas of interest from individuals in the field. Therefore, we undertook a series of conversations with policy actors, including researchers (scholars and other experts versed in quantitative evaluation techniques) and practitioners (policymakers, administrators, and program staff). To identify potential interviewees, we created a list of experienced policy actors that served in a variety of roles in the policy and research spheres. We then practiced purposeful sampling from this list to ensure an appropriate balance of participants and diversity of opinions. To facilitate the study and encourage candid conversation, we chose many of our interviewees based on our prior work and experiences (Maxwell, 2005). We sought researchers from a variety of disciplinary backgrounds to guard against the possibility that prevailing opinions from a certain field would unduly influence our findings. To ensure that our interviewees had an understanding of the challenges unique to educational research in general and postsecondary educational research in
particular, we made sure that all policy actors we interviewed had conducted work in or had some exposure to program evaluation in postsecondary education.  

To jumpstart these conversations, we distributed a seven-question online survey, which began by asking interviewees to rate their comfort with both producing and understanding evaluation research using non-experimental methods. This initial question helped us gauge the familiarity of the respondent with the methods and his or her experience working with evaluation. Next, we asked respondents to score various methods based on their perceived rigor, accessibility, ease of execution, and prevalence. Although a plethora of options exist, we focus on five commonly used methods in evaluation research in postsecondary education: (a) use of all non-participants; (b) exact univariate matching; (c) exact multivariate matching; (d) propensity scoring; and (e) regression discontinuity. We provided a brief description of each of the methods discussed in this study to facilitate conversations with individuals who were unfamiliar with one or more methods (see Appendix). The survey concluded with a question asking respondents to rate how much good quantitative evaluation research exists in postsecondary education today. In total, we received responses from all 16 individuals that received the survey.

Participants in this study come from a variety of professional backgrounds, including four faculty members, two university researchers, two regional educational organization staff, four state or federal government employees, one former legislator, one legislative chief of staff, and two college access program administrators. Many of the participants had worked at several organizations during their careers, spanning the local, university, state, and federal levels. Six participants gave themselves a four or higher (on a five-point scale) in terms of their comfort producing postsecondary education evaluations using non-experimental methods. Thirteen participants gave themselves a three or higher in

---

2 We set a goal of approximately 15 participants for this study. To reach this number, we contacted 24 individuals, of whom 16 agreed to participate, two declined, and six did not respond.

3 Although this approach differs from traditional comparison group construction methods, its emergence as a promising alternative, along with its acceptance by organizations like the Institute of Education Sciences, warrants its inclusion in the study.
terms of understanding these evaluations, with nine of these participants ranking themselves a four or five.

To help analyze the survey results, we also split participants into two categories: researchers and practitioners (described above). Of our 16 participants, we categorized eight as researchers and eight as practitioners based on their professional experience and technical training. Researchers averaged a 4.3 in understanding these evaluations and a 3.7 in comfort producing them. Practitioners, on the other hand, averaged a 2.7 in comfort understanding evaluations and a 1.7 in producing them.

In addition to the survey, we also examined articles on non-experimental evaluation methods in postsecondary education. Because policy actors use a variety of resources to locate relevant work, we searched many electronic databases and websites, including those that focus on academic work (e.g., JSTOR, Google Scholar, and the Education Resources Information Center), those that are targeted towards a more practitioner-based audience (e.g., the Pathways to College Network Online Library, the Pell Institute and Pathways to College Network Evaluation ToolKit, and the What Works Clearinghouse), and those designed to reach policymakers (e.g., National Governors Association’s Center for Best Practices, the Education Commission of the States’ Selected Research and Readings, and the National Conference of State Legislatures’ Document Library). We looked for trends in methods used and recommended for use in other studies. This information gave us a better sense of the postsecondary educational research landscape.

Shortly after respondents completed the survey, we engaged in follow-up conversations where we asked a series of open-ended questions informed by our database scan and interviewees’ survey responses. In particular, we inquired about how respondents chose their methods when conducting non-experimental evaluations and how they judged the methods chosen by others. We also asked whether it is better for researchers to prioritize rigor or accessibility and if there is a tradeoff between the two. We concluded by inviting respondents to describe how valuable this research is in creating
postsecondary educational policy. Although each interview followed the same basic structure, we allowed survey responses and the results of prior interviews to inform our approach with each interviewee and made adjustments when appropriate to ensure that this study would accomplish our intended goals (Maxwell, 2005). In total, we conducted 15 interviews.\(^4\)

Throughout all phases of this project, we adopt a pragmatist approach, focusing not on whether each method produces “correct” results but on the implications of using each method, how participants perceived these methods, and the usefulness of work using these methods to different groups (Creswell, 2007). We do not attempt to determine a “best” method or assume that a “best” method can possibly be defined, but we also do not assume that the usefulness of each method is entirely relativistic and dependent on the observer (Cornish & Gillespie, 2009). Instead, we focus on how researchers and practitioners describe each method and how non-experimental work can help solve practical problems in educational research (Crotty, 1998; Hookway, 2008).

**Themes**

We group our findings into four major themes: (a) the value and volume of non-experimental research, (b) the factors that impact selection and assessment of methods, (c) the perceptions of individual methods, and (d) the tension between rigor and accessibility.

**The Value and Volume of Non-Experimental Research**

Participants viewed non-experimental evaluation research as a valuable tool in creating good postsecondary educational policy. They used words like “tremendously,” “extraordinarily,” “greatly,” and “extremely” to describe the usefulness of this research, which also was perceived as “critical” in the policymaking process. One researcher stated, “Absolutely, I think evaluation research is critical to the

\(^4\) One survey respondent was unable to complete the interview in time to be included in our analysis. This participant was classified as a practitioner and is included in the survey results but not the interview analysis.
formation of good policy. I just in principle believe that. God, I hope that's true.” Participants argued that evaluation research “can really help drive decisions about allocation of resources in a really important way” and “allows you to have a conversation without people feeling personally attacked.” Another believed this research “sets a road map of which way to get started. I do believe you have to make adjustments continuously along the way, but no one ever got on the highway and took their hands off the wheel either.”

Most researchers recognized the need for an alternative to the experimental approach, even while praising randomized controlled trials. One participant stated, “it’s always great if you can do experimental research, but there simply will never be enough resources to do that for all the questions we need to answer.” Another expressed doubts about the financial and long-term research implications for experimental work, saying that “I'm not sure that it actually would be worth the money you would have to spend to exert such control in an educational setting...what may be most appropriate is to simply push education policy in the right direction.” For this participant, replication was key, and non-experimental methods allow for the replication necessary to making broad conclusions in statistical work because the costs are significantly lower.

Although there was near-universal agreement on the importance and value of non-experimental methods in postsecondary educational evaluation research, participants varied in their perception of how much good work exists today. Both researchers and practitioners averaged around a 2.8 (out of 5) when answering this question on the survey. But in the interviews, some respondents questioned how much of this research really qualifies as quality work. As one practitioner stated, “There's a lot of research out there. I'm just not sure how good it all is...I see a lot of holes in a lot that's out there and they're coming from really credible places. I'm not sure people are really scrutinizing it.” A researcher echoed this sentiment, arguing that
I am looking every day at evidence where people—people who normally do rigorous work—will do a very shoddy little pilot study and take it on the road. And I am so offended by it because I think it casts a shadow over all the rest of us and I think it does real damage... it's a professional crime... Because the strong likelihood is that if it fits the political agenda, they [policymakers] are going to act on that and you have just as much of a chance of hurting or cutting a program or putting money into a program that does not work as you do of accidentally doing some good.

Other participants reinforced this idea that both high-quality and low-quality work may have the potential to influence policy.

Other policy actors believed that no matter the amount of good research we currently have, we still need to create more to help better understand the opportunities and challenges ahead. One researcher argued that “we share a collective responsibility for the fact that we don't know more than we do about how to improve higher education practice.” Another believed that although a decent amount of research exists, much more is needed because

...the stakes are higher. Colleges and universities have gotten away with not actually turning the microscope inwards. We do a hell of a job with lots of other stuff about non-colleges and universities, but it’s about time that we apply the same approach here to policymaking and decisions about creating a new program or continuing an existing program. It’s got to be much more important.

Perhaps unsurprisingly, none of our participants indicated that they were completely satisfied with the amount and/or content of existing research.

In addition, our database scan revealed that a large volume of the research that does exist may be practically invisible to certain interested parties. For example, databases like JSTOR and Google Scholar, which focus on scholarly publications, contain a preponderance of articles using propensity scoring—many of which are stuck behind pay walls that make them essentially inaccessible to non-
academic audiences. Meanwhile, policy-focused databases like the National Governors’ Association’s Center for Best Practices and the National Conference of State Legislatures’ Document Library do not contain much evaluation work, instead focusing on summaries of bills and policies enacted across the United States. The What Works Clearinghouse, a practitioner-focused website established by the U.S. Department of Education, prioritizes randomized controlled trials above all other work. Simple Google searches yield studies using a great variety of methods, but it can be difficult to assess the quality of work found through these searches. Therefore, although quality work may be available through certain venues, it is not always available or easily accessible to all types of policy actors.

**The Factors That Impact Selection and Assessment of Methods**

Participants who conduct non-experimental evaluation research cited several key factors they consider when selecting methods. First and foremost, the research question guides their selection process. As one researcher argued, “define your question, make it falsifiable, and after that, look at the situation in front of you.” Other factors, such as data availability, expertise, cost, and time constraints, played into the choice. One researcher described how technical skills can help make the decision:

Sometimes when you want to do a really rigorous design but you are not completely there yet technically, you probably would rather go with a safer choice when you are 100% confident that you are producing the right results without limitations attached to it instead of producing something fancy but wrong.

Another researcher agreed, stating that

I am very sympathetic to picking a method that’s feasible and within your skill level, assuming the person is upfront about what they chose and why and it doesn't seem like they are intending to be biased. When it seems like they're intentionally choosing a group that's going to be doing worse, then I am offended, but short of that, I can be indifferent.
All of our participants described method selection as a process rather than a simple decision, and all seemed open to exploring more than one option instead of defaulting to a particular favorite.

Finally, several participants cited context as a critical factor in non-experimental method selection. To our participants, context encompassed many different things, including the location of the program being evaluated and the standards imposed by groups with which they worked. Interestingly, participants did not emphasize stakeholder needs as a key factor in methods selection. Although several researchers highlighted the importance of considering your audience, they did so mostly in regards to understanding their skill level at interpreting results and less as a driver in the methodological choices. It is possible that participants excluded this idea because these needs are folded into other factors, such as research question or audience.

Participants generally described using a process similar to that described above when evaluating the methods employed by other researchers. In addition, they cited the importance of having results that are appropriate for the intended audience and a suitable comparison group. One practitioner described the need for meaningful comparisons for policymakers to really feel like the evaluation is useful. As an example, he argued that although the University of California–Berkeley and the University of Wisconsin–Madison might be “academically on the same par,” the substantial differences between the California and Wisconsin contexts make it so that research performed in one location might not be applicable to the other. He stated that “When you hear that kind of comparison, it doesn't hold as much weight because it’s not apples-to-apples.”

Interestingly, some participants—both researchers and practitioners—were reluctant to judge the appropriateness of researchers’ method selection. Instead, they relied on knowledge from other sources and experts to help make those judgments. One participant described how he would “Google the test to learn more” if he encountered something with which he was unfamiliar. If he had to, he
would look at the methodology section, but “usually [his] eyes glaze over.” Another participant, a professional evaluator, explained why she does not necessarily judge the work done by others:

Most of us who work in higher education as faculty or staff are very critical. We’re kind of trained to be that way. So it’s really easy to go and read something and think, “Wow, they could have done this better.” And I certainly do that myself. But most things that are written, and even when I publish I think, “Gosh, I could have done that better”... But you get their best work at that moment...I don’t usually question the methodology. I might question their use of a tool or the statistic they ran or the numbers or something like that, but I don’t usually question the methodology. I figure they chose the best one they think would work to answer that question.

However, some other participants’ concerns about the overall quality of evaluation research appeared to be driven by methodology; this was particularly true among researchers.

No matter how participants selected methods or judged others’ choices, they universally called for others to engage in explicit discussion of assumptions, choices, and limitations. One researcher argued that we have to get over trying to be “right,” which seems to be the gist I see in a lot of academic journals, and just make sure it’s useful; and if you have to caveat it and say, this is useful up to this point, then it’s useful up to this point and beyond that, who knows.

A paragraph or two outlining the limitations of a study made one researcher “very forgiving,” especially when the authors “recognize that this is all [the research] can do.” Other researchers agreed that this explicit discussion helps create more powerful findings because it helps by “taking away the barriers to verify.” By making the explanation of decisions clear and complete, other readers “take you seriously.” One researcher suggested,

If there’s a limitation or decision you’ve made, make it explicit and justify it because in the end...if you try to hide the flaws or don’t explain why you made decisions, reviewers are going to
ask why didn’t you do this and why didn’t you do that? If you anticipate what people’s questions are going to be...you can head off those questions and people realize, oh, you did give thought to that. It’s more credible the more open and clear you are about the choices you made, and it also allows people to replicate it, if they can, in the future, which is what we want to do.

In addition to enhancing the replicability of a study, clearly stated limitations help create answers “because 9 times out of 10, if you can identify the problem and describe it, you can also begin to resolve it.”

**The Perceptions of Specific Methods**

We asked respondents to assess their familiarity with each of the five methods discussed above, as well as the rigor of each method. The survey revealed an overall inverse relationship between respondents’ assessments of familiarity and rigor. The use of all non-participants ranked as the most familiar method, followed by exact univariate matching, exact multivariate matching, propensity scoring, and regression discontinuity. Respondents’ rankings of rigor were the exact opposite, with propensity scoring and regression discontinuity seen as the most rigorous. In fact, one researcher commented on the high standing of these methods in the academic community by arguing that “you can put propensity score matching in your title and get accepted [for journal publication].”

Unsurprisingly, a similar inverse relationship emerged between respondents’ assessments of rigor and their assessments of how easy each method is to understand. For example, if a respondent ranked propensity scoring and regression discontinuity as highly rigorous, he or she would then score them low on ease of understanding. Respondents’ assessments of the ease of execution of each method followed the same trend as familiarity and ease of understanding, with the notable exception being that respondents judged propensity scoring, on average, as less difficult to execute than regression discontinuity.
When disaggregating findings into researcher and practitioner groups, results remained relatively consistent, with some interesting differences. The ascending/descending patterns mentioned above were roughly identical for the two groups, although practitioners expressed much lower levels of familiarity across all methods and somewhat higher opinions of the rigor of all methods. Surprisingly, practitioners described all methods except exact multivariate matching as slightly easier to understand than did researchers. However, another notable difference was a much greater prevalence of “not sure” responses among the practitioner group, which might indicate that practitioners were likely to answer some questions only if they understood the methods mentioned. This trend also played out during our interviews, as many practitioners were unwilling to discuss particular methods because they did not see themselves as qualified to do so.

Despite this ranking of rigor, many researchers stated in the interviews that no method is inherently more or less rigorous than another; instead, it depends on the question asked. As one researcher stated, “each method has its own value and there’s really no better or worse.” Another argued that there is “nothing wrong with any of these approaches...the question is, are they or are they not appropriate for the specific situation at hand?” Others echoed this sentiment; several respondents had trouble answering the survey questions about rigor because they were uncomfortable forcing themselves to evaluate rigor in generic terms. As one researcher stated, “rigor isn’t fundamentally in the methodology; it’s how the methodology is practiced that makes it rigorous.”

In general, practitioners wanted quality research that made sense to them. As one respondent described, “If it’s policy work, not pure academic research, [it should be] something your smarter-than-the-average-bear kind of person can grasp without necessarily having to be a psychometrician or an advanced statistician.” For this group, they were not necessarily aware of the details behind each method, and some were not particularly interested. Instead, they wanted to know that the work was
done well for the question at hand and trusted that the method chosen was best. As one practitioner stated,

I’ve looked at academic research and it’s just kind of amazing, the methods section and all their charts with the Rs and stats stuff, and that’s kind of irrelevant to people like me. You want to know that it’s a good study, well done, and that what they’re reporting is actually happening, that they can prove that it’s being shown.

The Tension Between Rigor and Accessibility

Before we began the interviews, we knew we wanted to address the perceived conflict between rigor and accessibility. The literature on the policy/research disconnect discusses this tension, where it is generally assumed that scholars prioritize rigorous methodologies while policymakers prefer studies that produce easily understood findings and/or minimal disruption of a program (Birnbaum, 2000; Weiss, Murphy-Graham, Petrosino, & Gandhi, 2008; Braverman & Arnold, 2008). Our database scan seemed to point toward this tradeoff: the perceived “rigorous” methods like propensity scoring and regression discontinuity tended to appear in less accessible venues focused on highly academic work but infrequently in more accessible places designed to reach policymakers and practitioners. In some cases, articles applying certain methods to particular topics are virtually unavailable to some groups that might be interested, so access to both the findings and methods is limited. In our work, we frequently struggle with this tension, and wanted to use this initial study to explore the issue.

In our interviews, we asked whether it is better to prioritize rigor or accessibility and about the tradeoffs, if any, between these two goals. This question prompted the most lively and lengthy discussions during our conversations. Not surprisingly, many participants struggled with choosing one over the other, and some outright rejected the idea that rigor and accessibility are conflicting goals.
Overall, most researchers and practitioners who answered this question called for an emphasis on rigor, regardless of how they chose to define the term. They argued that rigor should come first, and then accessibility should follow. As one participant stated, “It's got to be rigor, especially if it has the potential to influence programs or policy. If it’s an accessibility problem, that’s my problem as a consumer or as someone participating in an evaluation.” One researcher who is a faculty member claimed that “scholars, whether they admit or not...would take rigor in a heartbeat.” As one practitioner described, rigor “gives it the weight.”

Many respondents expressed a belief that rigor must be prioritized because authors can always make results more accessible. As one practitioner described, “I always would rather have the data be rigorous and comprehensive and precise because I can probably help interpret that or explain that to the end users in a layman's way.” And even if authors cannot make all pieces of the analysis accessible to a wider audience, some researchers felt this was not an issue. As one researcher described, “Make sure your rigor is there. Make sure your attention to detail and theoretical soundness is the appendix. Give them something that might require a leap of faith.” Others argued that even if you are able to reach fewer people, it has more power by being more rigorous. As one practitioner stated, “the audience might be a little bit smaller, but usually, you find out more.”

Some respondents argued that there is not a tradeoff between rigor and accessibility; in fact, evaluation research should strive to be rigorous and accessible, no matter what methods are chosen. For these respondents, the rigor came from the appropriate methodological choice, while the accessibility came from how those results are communicated—the “messaging” afterwards. Accessibility “comes in the dissemination and presentation, rather than the rigor.” As one researcher described, rigor doesn't have to come at the expense of accessibility, and vice versa. I think with the more recently popular designs, such as regression discontinuity and propensity score matching, it’s absolutely possible in my mind to make these very accessible information, even if the audience
doesn't have to know how to do it, doesn't have to understand all the statistical formula behind those, conceptually they can be accessible. I'm an idealist.

Some databases, like the What Works Clearinghouse, attempt to distill rigorous research down into clear takeaway points; others merely present the research in its original form, regardless of accessibility.

Although the importance of rigor drove our conversations, respondents still argued that concerted effort towards accessibility was crucial for evaluation research. One researcher argued that If your audience is someone who is not going to have the time or means to look up what you're doing and fully understand it, then it's more important to make it understandable to your audience, assuming you're not sacrificing the entire integrity of the project.

Respondents highlighted the jargon and technical details in academic publications, describing how this level of detail may not be necessary for many policymakers. As one researcher stated,

The simple fact of the matter is I have yet to meet a policymaker who enjoys thinking in terms of multivariate distributions and t-values and t-statistics. They don't. They simply don't. They want to know yes/no, up/down, viable/not viable, red/blue...that's what they're looking for.

Another questioned the usefulness of rigorous results that lacked accessibility, saying

There's face validity that's associated with an evaluative framework that can be communicated fairly broadly and so even though it may be strong and defensible purely on methodological grounds, it may not be something that can be communicated broadly and generate a lot of confidence with different stakeholder groups.

Although many of our participants were willing to prioritize rigor or accessibility, the general theme that emerged from our interviews was that the two were not mutually exclusive, and that the ideal study would involve at least some combination of sound methodology and clear descriptions of findings.
Moving Forward

We designed this study to take the first, small step towards a larger discussion. Evaluation is likely to become even more important in higher education in the coming years, and as such, those in the postsecondary educational community have a tremendous opportunity to make this research a meaningful part of the policymaking process. But to do so, policy actors need to engage in a transparent conversation about how we conduct evaluation, how we judge the work of others, and what these decisions may mean moving forward.

The limitations of this study are clear. The small sample size, combined with the sampling technique, limits the generalizability of the findings. We only examined non-experimental quantitative methods—of which we chose to highlight only five possible approaches—which means there are lots of areas within this field still left to cover. Further, the terminology used in non-experimental research frequently varies. The terms we used may mean different things to different people, so it is possible that our interpretation of non-experimental research methods could differ from those of our respondents. The survey instrument, which we designed to jumpstart our interviews and give some background information on perceptions, was left open-ended enough to create differing understandings among respondents. Although our participants worked in a variety of roles and organizations, the majority live and work in the Midwest where the policymaking environment could have unique characteristics. Taken together, these limitations circumscribe the generalizability of the findings. To further our understanding, we would need, at a minimum, more participants, more data, and more time.

However, we did not approach this study with the goal of developing generalizable results; instead, we aimed to investigate a crucial topic and generate points for future discussion. Therefore, despite these limitations, we believe that the information discovered in this study can help contribute to naturalistic generalizations among our audience (Stake & Trumbull, 1982; Cresswell, 2007) and point the postsecondary educational research community toward larger questions, such as those discussed below.
How do we empower those who consume, but do not necessarily produce, evaluation research to feel qualified to offer critiques? Are we comfortable asking policy actors to take a “leap of faith” when reading evaluation research?

Throughout our surveys and interviews, practitioners were consistently averse to discussing strengths and weaknesses of particular methods. Although we cannot expect individuals to offer opinions based on limited information, we were surprised by the degree of deference afforded to researchers. It appears that many consumers are willing to take an uncritical approach when using the work derived from certain methods; instead, they take a “leap of faith” and assume that the methods are appropriate and sound. As a result, the state of evaluation research may be such that a select group of researchers has the power to set research standards that may go unchallenged by the individuals who, in theory, should be using research results. Although specialization may be necessary to some extent, the benefit of a two-way exchange could help refine methodology in the future by calling attention to the realities of policymaking and practice.

This also raises the question of whether researchers need to figure out how to communicate their methods in a way that allows all types of consumers to understand all of the components of their work. Even if we assume that evaluation research must be made as accessible as possible, the question remains as to whether that burden should fall on researchers themselves. For a variety of reasons, many researchers might see their primary audience as the research community, not the broader community of policy actors; as such, accessibility may not be their chief concern. In these cases, perhaps the burden falls on other actors to step in to “translate” research findings.
When assessing evaluation research, do we—as a research community—really treat all methods as equally rigorous, depending on the question? Do publication trends—both in academic and non-academic venues—support this idea?

Many participants argued that every method could be rigorous, depending on the situation. However, it is difficult to reconcile these statements with the relative frequency of methods found in academic publications. Exact univariate matching may be appropriate in certain circumstances, but it is nearly impossible to find the method used in peer-reviewed journal articles; conversely, propensity scoring appears much more frequently than any other method and is applied to a wide variety of topics and questions. Propensity scoring in education is relatively new, so perhaps its novelty contributes to its popularity and perceived rigor. One participant spoke to this point, saying that “it’s never the case that there’s not going to be a new way to do things that people are going to claim is better, and it’s all guesswork anyways... even the most rigorous method is still going to be wrong because it’s all estimation.” Whether it is due to the questions we asked or overall trends in research methodology, some consideration must be paid to the fact that although researchers may say that all methods have value, in practice, we use them in differing amounts.

Have we, as a research community, worked to empower those outside our community to take on rigorous evaluations? Do we want to, and if so, how would we do that?

An enduring economic recession, a shift toward smaller government, and increasing accountability tied to programs and policies all contribute to an environment that calls for more evaluation efforts without accompanying increases in funding to pay for this research. There may be more evaluations that need to be done than there are trained evaluators and researchers to do them. This gap means that an increasing burden will fall on practitioners. In our interviews, we heard practitioners call for more help and training to allow them to do good evaluation work. One respondent,
for example, when asked what would be most useful from the research community, said, “Being able to take something that’s very academic and heavy and to be able to distill it down and to train community organizations or non-profits on how to do it well.” We should consider exploring new possibilities for either increasing the number of researchers working on evaluations or helping train practitioners to take on this work and do it well. By doing so, we can maximize the resources we have to find out more than we currently know.

In a time where publication venues are cutting back on space, how do researchers make their assumptions and limitations explicit?

Throughout our interviews, we heard researchers and practitioners call for more explicit discussion of assumptions, limitations, and methodological choices of authors in published work. At the same time, they recognized that authors may engage in these explicit discussions during the drafting stage, but they are later cut out by publishers and those releasing study results. Whether due to space constraints, financial concerns, or a belief that the audience would not want to see this information, these critical pieces of the puzzle seem to be getting lost despite continued calls for them. Most likely, this publishing trend will not reverse any time soon, so those producing this research must consider new ways to make assumptions and limitations explicit. Opening up the discussion of assumptions and limitations allows for more meaningful feedback from consumers, who also may be less likely to dismiss the findings from a study if they have a greater understanding of how important choices were made.

Are we reaching our intended audience with evaluation research? What do we need to do to expand our reach while still ensuring quality work is done?

Although the divide between researchers and policymakers was not the primary focus of this paper, many of the responses from our participants referenced this problem. It became clear through
our conversations that some types of evaluation research simply were not reaching practitioners and policymakers. Some participants spoke about ways that researchers could make sure that their studies had an impact. Several mentioned the importance of interpersonal interactions and being able to present a clear message. As one practitioner stated, “It’s not just the research but how it’s delivered, who it’s delivered to, and when it’s delivered.” It may be valuable for researchers to consider different ways of disseminating their findings. Our study only scratches the surface of the question of how to make academic research useful to a non-academic audience, and despite (or perhaps because of) the lack of an easy answer to this question, scholars must continue to investigate ways for research findings to become an integral part of the policymaking process, as well as to inform everyday practice.
References


Appendix: Method Descriptions for Survey and Interview

Evaluators use comparison groups to measure the behaviors, characteristics, and outcomes of participants in a particular program against a similar group of non-participants, which allows for a more accurate assessment of program effects. Included below are short summaries of five selected methods for creating a comparison group when using a non-experimental evaluation design.

**All non-participants** involves using all individuals that are part of the community from which participants are drawn but do not currently participate in the program. For example, a college access program within a particular high school might choose to compare participating students to all other students within the school.

**Exact univariate matching** involves creating a comparison group of individuals who match program participants based on one key characteristic. For example, an on-campus mentoring program focused on first-generation students could create a comparison group of all first-generation students on campus that are not in the program.

**Exact multivariate matching** involves creating a comparison group of individuals who match participating individuals on several key characteristics. For example, a program might create a comparison group of non-participating students that match participating students based on race, gender, and academic achievement prior to entering the program.

**Propensity scoring** uses regression analysis to predict an individual’s likelihood of participating in a program based on a set of key characteristics. After calculating an individual’s probability of involvement, program participants can be matched with non-participants who are equally or similarly likely to participate in the program. For example, a student enrolled in a tutoring program may be 86% likely to participate in the program based on his or her characteristics. For the sake of comparison, this individual then could be matched with the non-participant whose likelihood of involvement was closest to 86% or perhaps all non-participants with between 81% and 91% likelihood.

**Regression discontinuity** involves creating groups of participants and non-participants with characteristics that cluster around a cutoff point that determines enrollment in the program. For example, consider a merit scholarship that requires a 3.0 GPA. Regression discontinuity would compare students just below and just above the 3.0 GPA cutoff under the assumption that they are very similar, looking for discontinuities or irregularities around the participation cutoff point.