IMPROVING THE PRODUCTIVITY OF EDUCATION EXPERIMENTS: LESSONS FROM A RANDOMIZED STUDY OF NEED-BASED FINANCIAL AID

Douglas N. Harris
(corresponding author)
Educational Policy Studies
University of Wisconsin
at Madison
Madison, WI 53706
dnharris3@wisc.edu

Sara Goldrick-Rab
Educational Policy Studies
University of Wisconsin
at Madison
Madison, WI 53706
srab@education.wisc.edu.

Abstract

Given scarce resources for evaluation, we recommend that education researchers more frequently conduct comprehensive randomized trials that generate evidence on how, why, and under what conditions interventions succeed or fail in producing effects. Recent experience evaluating a randomized need-based financial aid intervention highlights some of our arguments and guides our outline of the circumstances under which the examination of mechanisms and heterogeneous impacts is particularly important. Comprehensive experiments can enhance research productivity by increasing the number of theories both tested and generated and can advance policy and practice by exploring the conditions under which interventions will be most successful in scale up. Paradoxically, while the emphasis on average treatment effects is typically associated with efficiency-minded economists, we argue that the approach is often inefficient from the standpoints of science and policy.

1. INTRODUCTION

In many ways, the advent of experimental design in the conduct of educational research represents scientific progress, focusing attention on whether interventions effectively influence outcomes. This type of evaluation is especially important for programs that expend considerable resources attempting to address significant public concerns. Too often, however, the manner in which educational experiments are conducted is inefficient from the standpoint of knowledge creation as well as policy and practice. A narrow scope emphasizing only the identification of average treatment effects (ATEs) fails to fully evaluate theories of change or provide a meaningful test of the degree to which policies are useful when brought to scale. Instead we recommend that researchers consider cost-effective ways to identify the conditions under which interventions are most effective and especially why they do or do not work. While the kinds of evaluations we describe are more expensive than usual, they may also generate a preponderance of useful knowledge in many cases, making them more cost-effective than typical experiments.

Need-based college financial aid is a good example of a policy benefiting from experimental evaluation, but one that could be inadequately assessed with the typically narrow evaluative focus. It is the most widely used and expensive intervention used to induce students to finish college—the federal government spends $155 billion on grants and loans each year (College Board 2010). A sizable body of rigorous research suggests that financial aid has modest positive effects on educational outcomes (e.g., Goldrick-Rab, Harris, and Trostel 2009; Deming and Dynarski 2010; Bettinger 2011). But estimates of financial aid's ATEs can appear small relative to their associated costs (Harris and Goldrick-Rab 2010), giving the impression that the policy may not be justified (Carey 2011). This potential consequence must be considered alongside strong political desires, given skyrocketing costs, to reevaluate the use of financial aid and consider ways to reform it in order to better meet the nation's education needs (Bettinger 2011). To guide reform agendas, researchers need to contribute a concrete and specific understanding of how financial aid changes student behaviors, under what conditions, and for which students. In this article we describe our recent efforts in that arena.

Understandably, the challenges of evaluating programs like financial aid, including the widely known problem of selection bias, have led researchers to prioritize the identification of internally valid ATEs (Shadish, Cook, and Campbell 2002). While a few studies have tried to consider variation in effects by specific student characteristics, many of the most rigorous studies either include fairly homogenous samples or rely on administrative data that include only limited student behavior measures. While there is also research on financial aid that includes larger populations and data from student surveys and/or

interviews, these are often not paired with rigorous experimental designs that allow for insights about student behavior to be connected to the estimation of treatment effects. And it is clearly difficult to account for an effect without first identifying the effect itself.

Beginning in 2008, we undertook the first randomized trial of a need-based financial grant. The decision to undertake this evaluation was the result of an uncommon opportunity—a privately funded program that used random assignment and accommodated a rigorous evaluation—and our evaluative approach was similarly unusual. Building on decades of research conducted by researchers in many disciplines, we constructed a robust data collection strategy grounded in competing hypotheses. Of course we planned to estimate the average effects of the program, but we also aimed to rigorously evaluate the suggestion made in prior studies that financial aid works differently for different students, under different circumstances, and for different reasons. Three years later, some of our expectations have been met: the program seems to be having heterogeneous effects, most of which we would have had difficulty identifying, let alone accounting for, had we not explicitly anticipated them.

We describe our evaluation strategies with the Wisconsin Scholars Grant (WSG) program in order to raise questions about the implicit assumption that evaluation approaches should be based first and foremost on establishing whether average treatment effects occur—before taking the necessary steps to identify heterogeneous impacts and mechanisms. We believe the usual approach increases the risk that researchers will overlook potentially important policy implications. For example, a program that is currently unsuccessful could be very successful if it were targeted or implemented in a different context. The strategy of engaging in robust data collection only *after* the identification of average treatment effects increases the likelihood that effect heterogeneity and mediators will never be explicated.

These are significant, widespread problems in education, and the research community is challenged to address them. As the past decade of experimentation associated with the Institute for Education Sciences has revealed, the effects identified in pilot studies are rarely replicated in other pilots or in scale up—and we rarely know why (Granger 2011; Schneider and McDonald 2006). Moreover, the pace of the scientific review process, and indeed the advance of time itself, makes it extremely difficult to collect the kind of information required to account for intervention effects after the intervention has concluded. Retrospective survey work and interviewing are hard to execute and are usually of lower quality (Mathiowetz, Brown, and Bound 2002). Understanding what works is important, but for the purposes of both scientific practice and policy it may be nearly as important to understand why and under what conditions programs *do not* work so that they can be improved and better targeted.

In this article we describe the results of our study thus far. In short, the intervention did not have the intended average effects, and if we had stopped there, as in the typical simple randomized control trial (RCT), we would have missed most of the story—about the heterogeneity, how the effects arose, and the reasons why the effects were not larger. We recognize that our approach was more costly than the typical RCT, but we believe the costs should be weighed against the benefits. Given the difficulty of quantifying the benefits of research, we instead propose a series of questions to guide researchers in deciding when to examine the mechanisms and heterogeneity of effects, or what we call a comprehensive experiment. These questions pertain, for example, to costs of collecting data on student characteristics and behaviors, and the likelihood that additional randomized trials might be possible in the future.

This rubric suggests that comprehensive RCTs are not justified in all cases but should be adopted more often than they are today, which is rarely. Below, we use our financial aid experiment to highlight what a comprehensive RCT can look like and to highlight the key considerations when deciding whether to conduct one. Other lessons from the experiment follow.

2. PRIOR THEORY AND EVIDENCE: THE BUILDING BLOCKS FOR RIGOROUS RESEARCH DESIGN

All scientific inquiry should be grounded in clear, thoughtful hypotheses. This is a strong protection against the intrusion of researcher biases and value judgments (Schrag 1989). Thus one of our initial steps in planning the randomized trial of need-based financial aid known as the Wisconsin Scholars Longitudinal Study (WSLS) was to write a book chapter in which we delineated competing theories about how financial aid might influence student behavior (Goldrick-Rab, Harris, and Trostel 2009).[1] Reflecting our respective backgrounds in sociology and economics, we developed an interdisciplinary perspective through detailed discussions, elaborations, and of course many drafts. We then laid a research design and data collection strategy onto that conceptual map. Here we elaborate briefly on each stage.

An Interdisciplinary Approach to Understanding Financial Aid

At its core, financial aid is a conditional cash transfer—theoretically, it provides income in exchange for behavior (college enrollment). Economic theory predicts that on the whole individuals will maximize their individual interests (utility) and respond rationally to incentives and resources, while

1. We are not the first to outline a variety of theories across multiple disciplines. See Cabrera, Stampen, and Hansen 1990; Paulsen and St. John 1997, 2002; Hossler, Schmit, and Vesper 1999; Berger 2000; Perna 2006; Dowd 2008.

sociologists emphasize the importance of groups and that responses will vary by the context in which the program operates and at whom it is aimed. These perspectives can be articulated in a set of specific theories and hypotheses grounded in key strands of thought within each discipline. Here we briefly summarize a few, to give a sense of how we used theory to develop a framework in which to design our study.

Economists typically theorize that investments in education (human capital) should work in much the same way as investments in physical capital—that is, individuals will invest in education insofar as the rewards exceed the costs and, specifically, to the point where the marginal benefits equal the marginal costs (Becker 1964; Manski and Wise 1983; Leslie and Brinkman 1988; DesJardins and Toutkoushian 2005). According to this assumption, people behave rationally and are well informed about their choices. While individuals might not be able to perfectly predict the costs and benefits of a long-term proposition like college, they are able to form unbiased expectations about what will happen under each scenario and can therefore make choices that maximize their present discounted value. Under these conditions a reduction in net price, whether through a reduction in tuition or an increase in grant aid or subsidized loans, should increase educational investments. Moreover, we should observe a straightforward average positive effect of financial aid, increasing college completion rates simply by reducing net costs.

An important implication of this simple version of the human capital model is that students from low-income families should enjoy just as much access to college as anyone else. By definition, they have smaller wealth endowments, but the basic model assumes that capital markets work perfectly so that students can borrow money to pay for college. Their incentives to invest are not smaller than those for students with larger endowments, since wealthier students forgo the interest earnings they would have accrued if they had invested their money in other ways rather than spend it on college. According to this approach, the interest rates received by the wealthy and paid by low-income borrowers are the same; therefore the total real cost of college education is also the same for both groups.

Why, then, are there persistent gaps in college completion rates according to family income? Economics can (though rarely does) address this by relaxing assumptions in the basic model, allowing for student-level variation with regard to returns to education (e.g., Carneiro and Heckman 2003; Cunha, Heckman, and Navarro 2005; Heckman, Lochner, and Todd 2006; Cunha and Heckman 2007), financial costs of education (Cameron and Taber 2004), "psychic" costs (Ehrenberg and Smith 2002), time horizon (Lawrence 1991), risk aversion (Hryshko, Luengo-Prado, and Sorensen 2011), and imperfect information (Ikenberry and Hartle 1998).

Behavioral economists, in contrast, question the assumption that people act rationally with regard to money (Thaler and Sunstein 2008). People are not only risk averse but more averse to losing what they already have, compared with missing the opportunity to add to what they have (Thaler 1981; Knetsch and Sinden 1984; Shefrin and Statman 1985; Odean 1998). This is known as loss aversion and might also be considered a bias toward the status quo. In the context of student loans, the problem is that students who take loans face a chance that they will be in a worse situation than they already are—that they will bear the cost of college without any benefit in earnings and will therefore lose some part of the lifestyle they already have.

Sociologists take a different approach to anticipating how individuals will respond to incentives and why that response might differ among individuals. In particular, sociologists draw attention to the role of context in situating individual decision making and behaviors and the role of institutions in circumscribing potential responses to incentives. An incentive administered to students in one context could affect them differently than an incentive administered in another context. For example, a scholarship might matter more to students attending resource-poor colleges, where having a scholarship is uncommon. Students might also respond more strongly to a grant if it is their only grant than if it were one of a jumble of several grants all with different rules (regardless of the total sum of the grants).

In addition, while economists view money as fungible—a dollar is a dollar—economic sociologists question the degree to which monetary incentives carry the same meaning for everyone (Zelizer 1994). Ethnographic evidence indicates that students make important distinctions among different kinds of money, with some forms striking them as particularly real, serious, and trustworthy (Clydesdale 2007). Thus effects of financial aid could vary based on how a student (or his or her family) understood the meaning of the money—and that meaning could vary if the grant were perceived as real (trustworthy) and stable rather than suspect, uncertain, and likely to disappear.

Both economics and sociology generate support for theories of how peer effects could shape the influence of financial aid. Drawing on theory and evidence from both disciplines, Harris (2010) explains a variety of both positive and negative peer influences and proposes a theory of "group-based contagion" in which people respond to others in the groups with which they identify themselves. If, for example, students identify with people who are "bad influences," the ties may make students worse off. This contagion model applies to college peers (e.g., Argys and Rees 2008) and to students who maintain close connections to friends from home, and it posits peer effects as moderating effects of financial aid. It could be extended further to consider alterations in social relations as a mechanism through which financial aid could

exert effects—for example, by making it more possible for students to live in dorms.

All these theories point to variation in financial aid effects, but each comes at the issue differently, and each suggests different mechanisms through which aid effects operate. We designed the WSLS to test these theories.

Evidence of Heterogeneous Effects

One reason we looked first at the theories about heterogeneity and mechanisms is that we sought to avoid and address the multiple comparisons problem. As the number of comparisons grows, so too does the likelihood that at least one difference will appear statistically significant, giving the possibly false impression of heterogeneity. We concur with Bloom and Michalopoulos (2010), who recommend having a theoretical justification and/or prior evidence of heterogeneity before conducting subgroup analyses. Below, we therefore complement our earlier discussion of theory with discussion of prior evidence of heterogeneous impacts of financial aid by program design, student characteristics, and contexts. We focus especially on rigorous studies of the effects of grants and scholarships on college entry, persistence, credits, and/or graduation.[2] We also adopt the convention that significant differences are those in which the impact coefficients are statistically different from one another across specific subgroups. Evidence is only "suggestive" if the individual coefficients are different from zero but not different from one another.[3] If a study reports results by subgroup but none of the coefficients is statistically different from zero, there is no heterogeneity.[4] Heterogeneity analysis inherently involves smaller samples and less statistical power; consequently there are many cases in which the point estimates are qualitatively different but the small samples yield wide confidence intervals.

Table 1 summarizes prior research according to several dimensions of heterogeneity by type of aid: need only (Kane 1995; Seftor and Turner 2002;

2. We specifically omit the following: (1) Linsenmeier, Rosen, and Rouse (2006) and van der Klaauw (2002) because they focus on attendance at a particular, anonymous university rather than any college enrollment; (2) some of the analyses within Kane (2004, 2007) that focus on effects of the Washington, DC, tuition program on college attendance in adjoining states; (3) Goodman (2008) because the effect is on public versus private college enrollment rather than any enrollment; (4) Dynarski (2004) because she focuses on college savings plans rather than grants, which both constitutes a different intervention and severely limits the sample characteristics because students with below-average family incomes have little opportunity to participate through savings; (5) Angrist, Oreopoulos, and Williams (2010) because they do not report effects on total credits or enrollment; and (6) Reyes (1995) and Dynarski (2005) because they focus on loans rather than grants and scholarships.
3. Our terminology here is similar to the distinction Bloom and Michalopoulos (2010) make between "suggestive" and "confirmatory" evidence.
4. This is not the same as a heterogeneity analysis, but these studies do still provide suggestive evidence. Given the difficulty of drawing conclusions about heterogeneity, we believe it is important to bring all the relevant evidence to bear.

Table 1. Number of Studies Finding Effects by Subgroup and Treatment Type

	Effect for Subgroup is:		
	Smaller/ Negative	**No Difference/ No Effect**	**Larger/ Positive**
Need only (Pell, Social Security)			
Women			1
Minorities		1	
Low-income/SES		1	1
Older/nontraditional			1
Low ACT/GPA			
Merit within need (GMS, Opening Doors, state programs)			
Women	1	1	2
Minorities		1	
Low-income/SES		1	3
Older/nontraditional			2
Low ACT/GPA			
Merit only (Canada STAR, state programs)			
Women		1	1
Minorities	1	1	1
Low-income/SES	1		1
Older/nontraditional			
Low ACT/GPA			2
General (GI Bill, tuition changes)			
Women			
Minorities		1	
Low-income/SES	0.5		1
Older/nontraditional			
Low ACT/GPA			

Notes: Numbers in the table represent counts of the number of studies with the specified heterogeneity. GI Bill studies counted as one-half because they are so far in the past, when the conditions of higher education were quite different. In the merit-within-need category, effects reported as "larger" if the program worked at all because by definition the programs are restricted to low-income students. Similarly, with Gates Millennium Scholars (GMS), because the sample is restricted to low-income minorities, these are cast as "smaller." See text for distinctions between suggestive versus confirmatory evidence in particular studies.

Bettinger 2004; Dynarski 2003), merit within need (Brock and Richburg-Hayes 2006; Barrow et al. 2010; Scrivener and Au 2007; Scrivener and Pih 2007; Kane 2003; DesJardins and McCall 2007), merit only (Dynarski 2000, 2008; Cornwell, Mustard, and Sridhar 2006; Scott-Clayton 2011), and general (essentially non-need, non-merit) (Angrist 1993; Stanley 2002; Bound and Turner 2002; Turner and Bound 2003). When we look across the four aid types we note that, despite the shift away from need-based aid, we actually know more about merit-based programs than about traditional need-based aid, for which we have only four rigorous studies. Three of the studies focus on the introduction and/or changes in Pell Grants that occurred many decades

ago. Given the changes in college costs in recent years and changes in the dynamics of poverty, it is not clear how informative these older studies can be in guiding present policy decisions.

From these studies, aid effects generally appear larger or more positive for disadvantaged groups. The only exceptions are the GI Bill, which was limited to males from fifty to seventy years ago, and Georgia HOPE, where researchers have come to opposite conclusions.[5] Our conclusion is somewhat different from Dynarski's (2002, p. 284) review, which concludes that "a given dollar of subsidy does not consistently have a larger impact on the schooling of low-income or minority individuals." The difference is largely explained by the number of new studies that have emerged in the decade since her review. It would be useful to consider heterogeneity along other dimensions, as we do ourselves later, though we are limited in this literature review by what these other studies have reported.

These conclusions, and the way we have approached our analysis of the WSG, also have important policy implications. Race and gender are often the focus of social science analyses because (a) these measures are typically available; (b) results so consistently vary along these dimensions; and (c) they are of considerable interest in the social sciences. However, they are not necessarily the most interesting from a policy standpoint because race and gender are often controversial as eligibility criteria for government programs. Income/socioeconomic status and academic ability are commonly used eligibility criteria, making the variation in impacts across these groups noteworthy. Financial aid programs, like many other areas of public policy, are blunt instruments (Bettinger 2011), but the very fact that not everyone is eligible for aid today suggests that targeting is possible.

Our review highlights several reasons why we designed and conducted the WSLS as we did. First, it points out the absence of randomized trials on the need-based aid programs that comprise the vast majority of aid dollars and that represent one of the single largest government social programs. In the process, we also provide the theoretical and empirical justification for heterogeneity analysis that avoids the multiple comparisons comparison—that is, we show that there are different mechanisms through which aid might operate and that there are good reasons to study heterogeneity along a range of dimensions, such as family income. We begin by introducing the grant program that is the subject of our analysis.

5. Dynarski (2000) found that the program increased the racial enrollment gaps, while Cornwell, Mustard, and Sridhar (2006) found that it reduced those gaps.

3. EMERGENCE AND DESIGN OF THE EXPERIMENT

The WSG is provided by the nonprofit Fund for Wisconsin Scholars, supported by a $175 million endowment created by John and Tashia Morgridge.[6] Decisions about program rules and operations are made by the fund, and we describe those most critical to this study.

The WSG provides university students with a $3,500 grant ($1,800 for two-year college students) per year for up to five years, making the total maximum award $17,500 per student. The grant is transferable among all public colleges and universities in Wisconsin, though it declines to $1,800 per year if students switch from a four-year to a two-year public college. (From this point forward, we focus only on the four-year sector.)

Students in this study were first notified that they would receive the grant during the second month of their first semester of college, specifically on 22 October 2008. Funds were distributed to their financial aid officers by the end of the term; for the vast majority of students the award appeared in their aid package in early December 2008. Subsequent payments arrived by the start of each new semester. Thus in their first year of college students received two grant payments (a total of $3,500) if they were eligible in both terms.

To remain eligible over time, the WSG rules required that students register for a full-time course load (12 credits) by the date of record. Students did not have to maintain continuous enrollment in order to receive the grant or earn specific credits during the term. The WSG was awarded at the start of each term and paid in full at that time. The fund did not remind students about eligibility criteria during their first year of college, though it did send emails in the second year. No performance level was required by the grant or mentioned by the program other than "satisfactory academic progress," which is required for all federal financial aid (and is typically, but not always, defined as a C average).

WSG students had to be Pell Grant recipients, Wisconsin residents, graduates of Wisconsin public high schools (GEDs allowed) during the three years preceding the commencement of college studies, enrolled full time in the first semester of college at any of the thirteen University of Wisconsin four-year institutions, and have at least $1 of unmet financial need. The program is not considered merit based, though students first had to apply to and gain admission to college and file a Free Application for Federal Student Aid (FAFSA) before qualifying for the WSG—substantial hurdles for many low-income students and their families (Dynarski and Scott-Clayton 2006a, 2006b). Given

6. For more information on the Fund for Wisconsin Scholars, see www.ffws.org.

this design, it is best to view our estimates as impacts on college outcomes conditional on college attendance, in contrast to prior studies that combine effects on attendance and persistence.

4. SAMPLE AND DATA COLLECTION

Next we provide an overview of the random assignment process, data collection, sample, and results. The full details are contained in a set of working papers found on our Web site (www.finaidstudy.org), including especially Goldrick-Rab et al. (2011).

In fall 2008, financial aid officers from the state's thirteen public universities identified more than three thousand new freshmen receiving Pell Grants as meeting the target criteria for the WSG. Per the fund's rules, the names of those eligible students were then placed in a database, from which six hundred were selected with assistance from the researchers using simple random assignment to receive the WSG (no blocking by campus was used in the random assignment). All six hundred students selected for treatment were immediately sent the award letter. From the remaining pool not chosen to receive the WSG, the researchers drew a stratified random sample of nine hundred students to serve as the control group.[7] The larger control group, combined with the oversampling of high-minority campuses, was designed to increase the possibility of identifying heterogeneous effects by race, informed by the theory and research noted earlier.

As a comprehensive randomized trial, we designed an extensive longitudinal data collection strategy, including annual collections of administrative data, lengthy surveys, and student interviews. With the exception of data from the National Student Clearinghouse and some de-identified financial aid application (FAFSA) data, all our data collection required obtaining consent from students. Wisconsin does not have a centralized data warehouse, and this required collecting data on students' financial aid packages and transcripts directly from each campus.

Data on financial aid packages were critical for understanding how the program was implemented and, in particular, how students experienced the grant. Because of aid packaging rules, we anticipated, and our research confirmed, that much of the grant was received in the form of loan reduction. The transcripts allowed us to estimate effects on many academic outcomes, including enrollment, credits attempted, credits completed, and grades. We anticipated variation in impacts on these outcomes based on the program's requirements for continued eligibility, although, as we show later, some of the impact patterns were not predicted.

7. We apply sampling weights as appropriate to address the oversampling.

Administrative data, as the name implies, typically focus on measures that are necessary to administer programs, but such measures are generally insufficient for testing theories about heterogeneous impacts of mechanisms. While income is one of the dimensions of predicted heterogeneity, the behavioral economic and sociological literature focuses on other factors. For this reason, we sent extensive twenty-six-page surveys to students three weeks after the treatment group was notified of the grant. The surveys included measures of risk aversion and family relationships that corresponded to the cross-disciplinary theories outlined earlier. These surveys obviously increased the cost of data collection and analysis, but we argue that this was worthwhile, given that this was the first randomized trial of need-based aid. To make the survey data collection as efficient as possible and to yield valid measures, we drew on the expertise of the University of Wisconsin Survey Center to, for example, reduce the cognitive complexity of survey wording and optimize the survey timing and use of incentives to attain maximum response rates.

While randomization facilitates internally valid impact estimates, this is insufficient for understanding the mechanisms of aid impacts. We therefore also conducted semi-structured interviews twice annually with a stratified random sample of thirty-six four-year students, allowing hypotheses to emerge and providing an alternative means for corroborating surprising or unclear findings arising from the quantitative data analysis.[8] Interviewers drew on these other data sources when developing questions and prompts with individual students. We then integrated the interview data with the survey and administrative data, creating a regular feedback loop for designing subsequent interview and survey instruments and facilitating triangulation in the analysis.

With our extensive data collection, we can examine sample characteristics along a variety of potentially important dimensions. Not surprisingly, given the eligibility criteria, 53 percent of students are first-generation college students, and their parents have income that is well below average, at about $30,000. While Wisconsin as a whole has a somewhat small minority population, the combination of our low-income population and our oversampling of high-minority campuses means that minorities are a much larger share of the sample (25 percent) than of the state as a whole.

8. Specifically, we selected four of thirteen universities and then drew a stratified random sample from about 50 percent of the full sample who consented to be interviewed; these were stratified based on treatment status, gender, and minority status (white or non–Southeast Asian versus any other racial/ethnic group). The selection of interviewees at random was important because it enhanced the generalizability of those results (Gibson and Duncan 2005). In this article we rely on the interviews as suggestive evidence about how students may have interpreted and/or used the treatment, and it is always triangulated with evidence obtained from surveys and administrative data.

5. OVERVIEW OF RESULTS

We have reported elsewhere an extensive analysis of the WSG experiment (Goldrick-Rab et al. 2011). Here we summarize the most salient findings from that study, some additional unpublished findings, and their relevance to our understanding of comprehensive experiments. All reported estimates of the grant's impact are based on intent-to-treat analyses.

Three years after the program began, our analyses revealed that the grant exhibited only a small positive impact, increasing credit completion rates for some students but failing to induce a statistically significant increase in any of the main academic outcomes we considered. A year and a half after the grant was awarded, treatment group students were 2.4 percentage points more likely to be enrolled, but this was not nearly statistically significant. Similarly, there were no significant effects on total credits or grade point average (GPA) and there were negative and significant effects on the percent of students reaching forty-eight or more credits by the spring of 2010. (This threshold is noteworthy because of the WSG's twelve credit hour threshold, which over four semesters adds to forty-eight credits.)

Did the program simply fail? Did students misunderstand it? Was the amount of money insufficient? One possible explanation, as indicated earlier, is that students were too uncertain about the program. Early interviews revealed that many grant recipients thought the program was a "scam." Given the apparent uncertainty in their minds, they may have simply ignored the WSG grant in their decision making and focused on information that was more certain (including other aspects of their aid package). Survey results also suggested they misunderstood the rules, believing that the WSG required a B average, which may have seemed out of reach. But even this misunderstanding would not explain negative effects around the forty eight credit hour threshold.

In addition, prior theory and research (see table 1), plus our own surveys and interviews, led us to posit that the grant was likely working better for some students than for others. We hypothesized that the effects would vary according to the overall level of advantage a student held as she or he entered college (prior to randomization). Those economic and social factors included an individual's academic preparation for college, family financial well-being, access to information, and familial resources and support. We combined these factors and estimated the propensity to persist in college over three years (using the control group) and then estimated the program impacts separately by probability tercile—the low tercile being the "unlikely graduates." This approach has also been used and discussed by others (Djebbari and Smith 2008), although it is not yet commonly used in education research.

Our results suggest that the modest positive average impacts are related to divergent effects operating for students in the bottom tercile compared

with those in the middle and top terciles. In fact, the students in the bottom tercile had large positive effects, while other students had negative impacts. We saw the same pattern of results in students' time use: those in the lowest propensity to persist group reduced their work hours and increased their study hours, but we did not see these responses in the top tercile (Benson and Goldrick-Rab 2011; Harris, Goldrick-Rab, and Taber 2011). It is no surprise, then, that the averages of these opposite effects, discussed earlier, were very small and statistically insignificant.

But why? How could giving students more financial aid drive down their persistence? It would not have been possible to develop a compelling answer to these questions without the combination of administrative records, surveys, and interviews. A closer examination of students' financial aid records, their high school coursework, their responses to survey questions about grant requirements and time usage, and their discussions of decision-making approaches about credit load all pointed in the same direction: the students who benefited from the WSG had weaker academic preparation and used the money to work less and study more. In contrast, students with stronger preparation were also receiving a federal grant known as the Academic Competitiveness Grant (ACG), which required them to earn a B average or risk losing their grant. Students were very aware of the ACG, more so than the WSG (which requires full-time enrollment) (Kinsley and Goldrick-Rab 2011).

One explanation we considered is that students prioritize academic achievement over credit loads—that they were trying to keep their ACG and while doing so risked losing their WSG. Given the byzantine process of aid packaging, and evidence that students forget about what aid they have (Angrist, Lang, and Oreopoulos 2009), this seems plausible. However, with extensive data collection, we were able to largely rule out this explanation.

We are finding more support for an alternative hypothesis: that ACG students were more likely to lose their Pell grants and this in turn triggered a loss of the Pell and WSG (recall that Pell eligibility is an eligibility requirement for the WSG). One possible reason for such a loss is that the ACG students were closer to the income threshold for Pell eligibility; an increase in income in the second or later years could push these students out of the Pell income range. Whatever the cause, the loss of aid gradually seems to have pulled the total grant and scholarship aid close to the level of the control group. If some students are loss averse, as we hypothesized in our earlier paper (Goldrick-Rab, Harris, and Trostel 2009), then the treatment may have actually had a negative impact on ACG students. Given the ACG receipt is correlated with the propensity scores, this could also explain the initial effect heterogeneity. Similar patterns emerge for other academic outcomes, as well as for how students used their time. Extensive sensitivity analyses confirm that these

heterogeneous effects are not caused by baseline nonequivalences or missing data.[9]

The development and testing of our hypotheses obviously required multiple theories and data sources. Without that, we surely would not have discovered either the possible heterogeneity or its possible causes, which have important lessons for the WSG program funders and policy makers.

6. LESSONS LEARNED

The experience of conducting a randomized evaluation of an intervention is a formative one for any social scientist, and especially for education researchers who have long struggled to find ways to establish the case for causal inference in their field. The time-intensive, intricate nature of the work and the challenges in documenting the interventions' components and tracing their effects have all been written about elsewhere, and we believe those lessons resonate here as well. What we hope to add to the discussion is consideration of ways to efficiently generate scientific and policy-relevant information from each randomized trial.

Criteria for Determining the Scope of an Experiment

The experiences we described helped us formulate a set of questions we think could be considered by funders and reviewers when assessing the value of research proposals, as well as by researchers when formulating those proposals. Taken as a whole, these questions are intended to weigh the costs and benefits of comprehensive RCTs. Our logic here is consistent with the approaches of some foundations (e.g., the William T. Grant Foundation) and some federal agencies (e.g., the National Institutes of Health) and is related to the idea of optimal design in which the costs of increasing the sample size are explicitly weighed against the additional statistical power. While we take what is in spirit an economic cost-benefit analysis, we argue that, in their justifiable quests for internal validity, economists and others are sometimes inefficiently focusing on average treatment effects.

(1) *Are similar interventions already using considerable public resources?* The Institute for Education Sciences guidelines discuss the related concept of "wide use," but it might be better defined in terms of "large resources." Some programs are widely used but involve few resources. Conversely, some programs might be used in only a few states but be extremely

9. We conducted F-tests within the ACG and non-ACG categories as well as a second test in which we estimated predicted persistence rates. In neither case did we find significant baseline differences between the control and treatment groups.

expensive. When considerable resources are at stake, the argument for a comprehensive RCT is stronger.

(2) *Are experiments on the topic rare and likely to remain so in the future?* If so, is it relatively difficult to conduct experiments on this topic? A variety of ethical, political, and feasibility challenges stand in the way of randomized trials in general, but the problems are greater in some areas than in others. As we discuss later, the challenges are particularly severe with need-based financial aid programs.

(3) *Does prior theory and/or evidence anticipate variation in program effects across designs, subgroups, and/or contexts?* This question should be addressed by considering a broad range of social science research, not only studies published in education. We say this because too often it seems that education researchers limit themselves to lessons learned only from that single institution (the school) when in fact educational interventions touch many aspects of students' lives (including the family, the church, etc.), so a wider range of lessons should be considered. At the same time, tenets for strong methodological practice must be employed, such that the researcher does not simply have a license to mine data for salient findings. In the WSLS, we started by writing a chapter on the theories about financial aid effects and reviewing evidence of heterogeneity (see table 1) and indeed found both theory and evidence supporting heterogeneity along specific dimensions.

Analogously, do we have much theory and evidence about the mechanisms through which the program works? To justify the comprehensive RCT, it is important to have one or more theories. If those theories have not yet been tested, the need for a comprehensive RCT is again greater.

(4) *Can data for analyses of mechanisms and heterogeneity be collected in a cost-effective manner?* This is perhaps one of the most difficult questions to answer, as the standard for cost-effectiveness depends on both the associated costs and the perceived value of the knowledge generated. However, we would note that if either a heterogeneity or mechanism analysis is called for, the marginal cost of adding the other is relatively modest. With either heterogeneity or mechanism analyses, some type of researcher-designed survey is likely to be necessary, and the cost of adding questions to a survey is low.

(5) *Do the researchers demonstrate a commitment to robust data collection by participating in an interdisciplinary team with knowledge of competing theories and methods?* There are almost always multiple theories of action, cutting across disciplines. It is therefore necessary for the research team and analyses to do the same.

(6) *Are there other potential uses of the additional data that will be collected?* Is there broader scientific merit to the analysis beyond the evaluation per

se? If so, and if the researchers seem intent on using the data for these purposes, more costly data collection might be worthwhile. The challenge here is that the standard for evidence on the associated analyses might vary; it is likely that the data could be used for descriptive purposes rather than evaluative ones.

The more frequently and confidently we can answer "yes" to the questions posed above, the stronger the case for a comprehensive experiment. Now we return to our financial aid experiment as an illustration and example.

Applying the Criteria to Financial Aid and the WSLS

By the way we described the Wisconsin Scholars Longitudinal Study in this article, it is clear that we believe the answer to most of these questions is yes. We began by describing the dominance of financial aid policies in the landscape of higher education and the growing expense of those policies. We also described the need for an experiment and reviewed both prior theory and evidence that strongly suggested treatment effects would be heterogeneous and the mechanisms through which they arose would be potentially numerous. We then described a strategy for data collection that was robust, supported by an interdisciplinary research team with experience in several methodologies, and would lead to the creation of a longitudinal database on low-income college students that could be leveraged for other research, such as in the area of college student time use, which has received considerable attention of late (Arum and Roksa 2011; Babcock and Marks 2011). While we recognize that our analysis is just getting started and that the costs of data collection and analysis have been substantial, we believe over time a strong case can be made that our overall approach was cost-effective.

The case is even stronger when we consider the ethical dilemmas and associated low likelihood that more experiments with need-based aid will occur in the future. With financial aid, we argue that the usual ethical dilemmas are more pronounced. Students do not usually have strong preferences to receive most (nonfinancial) higher education interventions, such as additional tutoring or small learning communities. Even when students are paid cash to receive tutoring, many do not take advantage of the program (Angrist, Lang, and Oreopoulos 2009). With financial grants the intervention is money, which is something that almost everyone wants and benefits from in some sense.

Randomized trials of need-based aid also pose even greater ethical dilemmas compared with other conditional cash transfers because (1) need-based aid is by definition aimed at low-income students who not only want the money but arguably need it to meet even the most basic expenses and (2) financial aid

involves large sums of money—up to $17,500 per student in the program we study. Most conditional cash transfers involve much smaller sums of money.

Finally, randomized trials are often justified on the basis of resource scarcity: there is often simply no way to provide the intervention to everyone. Thus, for example, there may be only so much money available to implement a whole school reform such as Success for All,[10] but such interventions are necessarily composed of a package of interventions. This is not the case with financial aid, which is easily divisible into smaller parts or dollar amounts. The program funders could simply reduce the amount of the grant for each student and give some small amount of money to everyone.

This is not to say an RCT is unethical. On the contrary, ethics would also seem to demand a randomized trial to demonstrate the efficacy of such a large resource outlay, and in this case the problem is largely addressed by the natural experiment created by the Fund for Wisconsin Scholars. Nevertheless, this yields a paradox: a program where the need for an RCT is extremely high is also one that is ethically, and consequently politically, most problematic. It is certainly possible that a similar experiment will occur in the future, but given the difficulties and the fact that decades have passed since Pell was instituted without one, there is not much reason to expect a similar experiment anytime soon. This, combined with the vast resources going into financial aid, would seem to shift the calculus clearly in the direction of a comprehensive experiment.

Methodological Challenges of Expanding the Scope

In addition to the broad choice between simple and comprehensive RCTs, we also believe there are lessons here for specific aspects of the analysis of heterogeneity and mechanisms.

Studying Heterogeneity

While we view heterogeneity as an important topic that deserves more attention (see question 3 above), it does not come without risks. When estimating impacts for multiple subgroups, the odds of a type I error increase (the "multiple comparisons" problem).[11] In one excellent overview on the topic, Bloom and Michalopoulos (2010, pp. 1–2) describe the conditions under which they believe this type of analysis is acceptable:[12]

10. See www.successforall.org.
11. The multiple comparisons problem is usually defined to apply to subgroup (heterogeneity) analyses as well as for different outcomes. The problem is somewhat different with subgroups because this always involves splitting the sample into smaller groups, which reduces the probability that any coefficient will be significant. With multiple outcomes, the sample size is often unchanged.
12. We omit one condition from this list: "Statistical significance of the subgroup's estimated intervention effect." This is omitted for brevity as well as because it will very rarely be possible to meet the second condition in the text if this is not also satisfied. The second condition in our list combines the two conditions into one.

- "Pre-specification of the subgroup" before conducting the analysis, based on prior theory and evidence.
- "Statistical significance of the subgroup's estimated intervention effect" versus some other group.
- Statistical significance of the ATE for the study sample.
- Presence of a consistent pattern or story.

There are several challenges in satisfying these conditions. For example, several factors drive down the statistical power in heterogeneity analysis. First, the standard ATE test is whether a coefficient is different from zero, which involves sampling error only in the ATE estimate, whereas comparing whether two coefficients are different involves sampling error in each. (Interestingly, researchers often do not carry out tests of whether point estimates are statistically different from one another and instead rely on informal judgments.) Second, the sampling error in each coefficient is larger than in the ATE analysis because of splitting the sample (the best-case scenario is a 50–50 split). In short, we are trying to estimate a smaller effect using considerably less statistical power. (With large administrative systems, statistical power is often a trivial issue, but recall that the measures available in administrative data systems are not the ones associated with hypothesized variation in impacts.) Notwithstanding the power problems, a few of the differences we observe are statistically significant across subgroups; given that this is part of a broad, consistent pattern of results (see Bloom and Michalopoulos's (2010) fourth condition), we find them compelling.

We also encountered a disciplinary divide as we presented this heterogeneity work to different audiences. In contrast to those concerned with the multiple comparisons problem, many economists argued that there must be students who are close to or "on the margin" of continuing in college, and the effect should be largest for these students. Some put this differently, arguing for a data mining approach, though the idea was basically the same. Whatever we might call it, this approach downplays the above four conditions and amounts to asking, For whom is the effect largest?

As we followed all but the third condition, our approach is somewhere in between data mining and the Bloom and Michalopoulos (2010) recommendations. While we found statistically significant average impacts on a handful of outcomes, we nevertheless conducted the heterogeneity analysis on all of them. We argue that as long as the other three conditions are met, analysis of heterogeneity is not only justifiable but very important. There is no reason to expect any relationship between the likelihood of a significant average treatment effect and the likelihood of heterogeneity. As our analysis highlights, heterogeneity can occur just as easily when

the ATEs are null. We were no more likely to find heterogeneous effects for the outcomes where we found positive ATEs. Given the clear policy implications of heterogeneity analysis for policy, we argue for a hybrid approach that relaxes the requirement of statistical significance of the average effects.

Mechanisms

A different set of problems arises in analysis of mechanisms. From the usual standpoint of causal inference, identifying the mechanisms behind programs requires solving the same problem as identifying the ATE on the main outcomes. That is, we have to solve the selection bias problem. The effect of the treatment on the mediator—the "first arrow"—is straightforward because we can still rely on random assignment to the treatment. However, the effect of the mediator on the main outcome (e.g., college enrollment)—the second arrow— is more challenging because participants are not randomly assigned to the mediator.

While the idea of using assignment to treatment as an instrumental variable (IV) to learn about the effects of mediators on outcomes has theoretical promise, it is generally impractical: IV requires larger samples than experiments typically allow, and there are generally multiple potential mechanisms, invalidating the IV assumptions.

Raudenbush (2011) provides a recent and very thoughtful discussion of the approaches to studying mechanisms. We agree with Raudenbush that there are ways to improve on the study of mechanisms, though these require strong assumptions, and we believe our study of the WSG highlights some additional alternatives. In particular, we argue for a strategy of triangulation across multiple data sources and mixed methods (Caracelli and Greene 1993). Social scientists have long argued for this approach, which has been undertaken in noted experiments such as Moving to Opportunity and New Hope (DeLuca and Rosenblatt 2010; Yoshikawa et al. 2008; Yoshikawa, Weisner, and Lowe 2009). In the WSLS we designed interview protocols with mechanisms in mind and then adapted these over time to test new hypotheses. Of course, we cannot subject interviews to the same types of statistical tests as survey and administrative data, and some argue that we would not want to (Small 2009). But given that the usual statistical standards are very difficult to satisfy in analysis of mechanisms—even with standard quantitative analyses—we and many others have argued that qualitative data analysis provides important complementary evidence (Gibson and Duncan 2005). Mixed methods can be used to test the assumptions required for quantitative analysis described by Raudenbush and to provide separate tests that in combination yield more convincing conclusions.

7. FINAL THOUGHTS

The design and implementation of randomized trials is not straightforward and involves judgment calls along a variety of dimensions. What we have attempted to do here is to highlight one aspect of these decisions that receives little attention: the comprehensive RCT versus the far more common simple RCT. By considering the costs and benefits of each approach, we follow in the footsteps of those who in years past used the same logic to justify experiments over quasi-experiments. As Cook (2002, pp. 176–77) writes, "Experiments are probably less expensive in the long run because, being more efficient about reducing causal uncertainty, fewer of them are needed for the same degree of confidence in the causal conclusion drawn." Similarly, we argue that if the goals are the generation of scientific and policy-relevant knowledge, the comprehensive RCT is also more cost-effective than the simple RCT in many circumstances.

The WSLS helps to illustrate our arguments. While there are no doubt specific things we could have done better, we believe the large existing public resources, the absence of prior RCTs, and other factors aligned to provide a strong case for our general comprehensive approach. The early returns seem to reinforce the point. First and foremost, if we had conducted an experiment of more limited scope, we might have reached very different conclusions. Without the FAFSA and survey measures, for example, one of the more plausible explanations for the heterogeneous effects would have been baseline differentials between control and treatment groups for one or more terciles—something we were able to partly rule out with our rich data collection. Further, had we not become "anthropologists of our own study," as recommended by Cook (2002), we would have had limited information about program implementation, making it much harder to explain the very modest ATEs. Once we consider how students received the money, it is perhaps surprising that any subgroup benefited in measurable ways.

Most important, the differences in results between simple and comprehensive RCTs also have real policy implications. If we had estimated only ATEs, the simple story would apparently have been "financial aid has no effect." Instead, corroborating prior quasi-experimental work, our analysis suggests that financial aid works extremely well for some students, perhaps by inducing them to study more and work less. As we move forward in our analyses, we will continue to look for ways that will not only allow us to better target aid but to better design aid programs to increase benefits for everyone. Given the $155 billion annual national investment, we believe the fraction of that amount we have spent for this comprehensive experiment is a small price to pay.

The study we discuss was made possible by generous financial support from the Bill and Melinda Gates Foundation, William T. Grant Foundation, Spencer Foundation,

Institute for Research on Poverty, Wisconsin Center for the Advancement of Postsecondary Education, and an anonymous donor. The research was conducted with the permission of the Fund for Wisconsin Scholars. We thank the staff and advisory board of the Wisconsin Scholars Longitudinal Study for their substantial assistance, and Howard Bloom and one anonymous reviewer for their useful comments. All opinions expressed are those of the authors, as are all mistakes.

REFERENCES

Angrist, Joshua D. 1993. The effect of veterans benefits on education and earnings. *Industrial and Labor Relations Review* 46(4): 637–52.

Angrist, Joshua D., Daniel Lang, and Philip Oreopoulos. 2009. Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics* 1(1): 136–63.

Angrist, Joshua D., Philip Oreopoulous, and Tyler Williams. 2010. When opportunity knocks, who answers? New evidence on College Achievement Awards. NBER Working Paper No. 16643. Cambridge, MA: National Bureau of Economic Research.

Argys, Laura, and Daniel Rees. 2008. Searching for peer group effects: A test of the contagion hypothesis. *Review of Economics and Statistics* 90(3): 442–58.

Arum, Richard, and Josipa Roksa. 2011. *Academically adrift.* Chicago: University of Chicago Press.

Babcock, Philip, and Mindy Marks. 2011. The falling time cost of college: Evidence from half a century of time use data. *Review of Economics and Statistics* 93(2): 468–78.

Barrow, Lisa, Thomas Brock, Lashawn Richburg-Hayes, and Cecilia Elena Rouse. 2010. Paying for performance: The educational impacts of a community college scholarship program for low-income adults. Federal Reserve Bank of Chicago Working Paper No. 2009–13.

Becker, Gary. 1964. *Human capital.* New York: Columbia University Press.

Benson, James, and Sara Goldrick-Rab. 2011. Putting college first: How social and financial capital impact labor market participation among low-income undergraduates. Unpublished paper, University of Wisconsin at Madison.

Berger, J. B. 2000. Optimizing capital, social reproduction, and undergraduate persistence: A sociological perspective. In *Reworking the student departure puzzle*, edited by John M. Braxton, pp. 95–124. Nashville, TN: Vanderbilt University Press.

Bettinger, Eric. 2004. How financial aid affects persistence. In *College choices: The economics of where to go, when to go, and how to pay for it*, edited by Caroline Hoxby, pp. 207–38. Chicago: University of Chicago Press.

Bettinger, Eric. 2011. Financial aid: A blunt instrument for increasing degree attainment. Washington, DC: American Enterprise Institute.

Bloom, Howard S., and Charles Michalopoulos. 2010. When is the story in the subgroups? Strategies for interpreting and reporting intervention effects on subgroups. New York: MDRC.

Bound, John, and Sarah Turner. 2002. Going to war and going to college: Did World War II and the G.I. Bill increase educational attainment for returning veterans? *Journal of Labor Economics* 20(4): 784–815.

Brock, Thomas, and Lashawn Richburg-Hayes. 2006. Paying for persistence: Early results of a Louisiana scholarship program for low-income parents attending community college. New York: MDRC.

Cabrera, Alberto F., Jay O. Stampen, and Lee W. Hansen. 1990. Exploring the effects of ability to pay on persistence in college. *Review of Higher Education* 13(3): 303–36.

Cameron, Stephen V., and Christopher Taber. 2004. Estimation of educational borrowing constraints using returns to schooling. *Journal of Political Economy* 112(1): 132–82.

Caracelli, V. J., and J. C. Greene. 1993. Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis* 15(2): 195–207.

Carey, Kevin. 2011. $44-billion ought to buy some accountability on campuses. *Chronicle of Higher Education*, 20 March.

Carneiro, Pedro, and James J. Heckman. 2003. Human capital policy. In *Inequality in America: What role for human capital policies?* edited by James J. Heckman and Alan B. Krueger, pp. 77–239. Cambridge, MA: MIT Press.

Clydesdale, Timothy. 2007. *The first year out: Understanding American teens after high school*. Chicago: University of Chicago Press.

College Board. 2010. *Trends in student aid 2010*. Available http://trends.collegeboard.org/downloads/archives/SA_2010.pdf. Accessed 1 December 2011.

Cook, Thomas D. 2002. Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis* 24(3): 175–99.

Cornwell, Christopher, David Mustard, and Deepa Sridhar. 2006. The enrollment effects of merit-based financial aid: Evidence from Georgia's HOPE scholarship. *Journal of Labor Economics* 24(4): 761–86.

Cunha, Flavio, and James J. Heckman. 2007. Identifying and estimating the distributions of ex post and ex ante returns to schooling. *Labour Economics* 14(6): 870–93.

Cunha, Flavio, James J. Heckman, and Salvadore Navarro. 2005. Separating uncertainty from heterogeneity in life cycle earnings. *Oxford Economic Papers* 57(2): 191–261.

DeLuca, Stefanie, and Peter Rosenblatt. 2010. Does moving to better neighborhoods lead to better schooling opportunities? Parental school choice in an experimental housing voucher program. *Teachers College Record* 112(5): 1443–91.

Deming, David, and Susan Dynarski. 2010. Into college, out of poverty? Policies to increase the postsecondary attainment of the poor. In *Targeting investments in children: Fighting poverty when resources are limited*, edited by Philip Levin and David Zimmerman, pp. 283–302. Chicago: University of Chicago Press.

DesJardins, Stephen L., and Brian P. McCall. 2007. The impact of the Gates Millennium Scholars Program on selected outcomes of low-income minority students: A regression discontinuity analysis. Working Paper, University of Michigan.

DesJardins, Stephen L., and Robert Toutkoushian. 2005. Are students really rational? The development of rational thought and its application to student choice. In *Higher education: Handbook of theory and research*, Vol. 20, edited by John C. Smart, pp. 191–240. Dordrecht: Kluwer.

Djebbari, Habiba, and Jeffrey Smith. 2008. Heterogeneous impacts in PROGRESA. IZA Discussion Paper No. 3362, Institute for the Study of Labor.

Dowd, Alicia. 2008. Dynamic interactions and intersubjectivity: Challenges to causal modeling in studies of college student debt. *Review of Educational Research* 78(2): 232–59.

Dynarski, Susan. 2000. Hope for whom? Financial aid for the middle class and its impact on college attendance. *National Tax Journal* 53(3): 629–61.

Dynarski, Susan. 2002. The behavioral and distributional implications of aid for college. *American Economic Review* 92(2): 279–85.

Dynarski, Susan. 2003. Does aid matter? Measuring the effect of student aid on college attendance and completion. *American Economic Review* 93(1): 279–88.

Dynarski, Susan. 2004. Who benefits from the college saving incentives? Income, educational expectations and the value of the 529 and Coverdell. *National Tax Journal* 57(2): 359–83.

Dynarski, Susan. 2005. Loans, liquidity and schooling decisions. Unpublished paper, Harvard University.

Dynarski, Susan. 2008. Building the stock of college-educated labor. *Journal of Human Resources* 43(3): 576–610.

Dynarski, Susan, and Judith Scott-Clayton. 2006a. Simplify and focus the education tax incentives. *Tax Notes* 111: 1290–92.

Dynarski, Susan, and Judith Scott-Clayton. 2006b. The cost of complexity in federal student aid: Lessons from optimal tax theory and behavioral economics. *National Tax Journal* 59(2): 319–56.

Ehrenberg, Ronald G., and Robert Smith. 2002. *Modern labor economics: Theory and public policy*, 7th ed. New York: Addison-Wesley.

Gibson, Christina M., and Greg J. Duncan. 2005. Qualitative/quantitative synergies in a random-assignment program evaluation. In *Discovering successful pathways in children's development: New methods in the study of childhood and family life*, edited by Thomas Weisner, pp. 283–315. Chicago: University of Chicago Press.

Goldrick-Rab, Sara, Douglas N. Harris, James Benson, and Robert Kelchen. 2011. Conditional cash transfers and college persistence: Evidence from a randomized need-based grant program. Institute for Research on Poverty Discussion Paper No. 1393–11.

Goldrick-Rab, Sara, Douglas N. Harris, and Philip Trostel. 2009. Why financial aid matters (or does not) for college success: Toward a new interdisciplinary perspective. In *Higher education: Handbook of theory and research*, edited by John C. Smart, pp. 1–46. New York: Springer.

Goodman, Joshua. 2008. Who merits financial aid? Massachusetts' Adams Scholarship. *Journal of Public Economics* 92(10–11): 2121–31.

Granger, Robert C. 2011. The big why: A learning agenda for the scale-up movement. *Pathways* (Winter): 28–32.

Harris, Douglas. 2010. How do school peers influence student educational outcomes? Theory and evidence from economics and other social sciences. *Teachers College Record* 112(4): 1163–97.

Harris, Douglas, and Sara Goldrick-Rab. 2010. The (un)productivity of American higher education: From "cost disease" to cost-effectiveness. LaFollette School Working Paper No. 2010–023.

Harris, Douglas, Sara Goldrick-Rab, and Christopher Taber. 2011. The causal effects of financial aid on time use among low-income university students. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness. Washington, DC, September.

Heckman, James J., Lance J. Lochner, and Petra E. Todd. 2006. Earnings functions, rates of return, and treatment effects. In *Handbook of the economics of education*, Vol. 1, edited by Eric Hanushek and Finnis Welch, pp. 307–458. Amsterdam: Elsevier.

Hossler, Donald, Jack Schmit, and Nick Vesper. 1999. *Going to college: How social, economic, and educational factors influence the decisions students make*. Baltimore, MD: Johns Hopkins University Press.

Hryshko, Dmytro, Maria Jose Luengo-Prado, and Bent Sorensen. 2011. Childhood determinants of risk aversion: The long shadow of compulsory education. *Quantitative Economics* 2(1): 37–72.

Ikenberry, Stanley O., and Terry W. Hartle. 1998. *Too little knowledge is a dangerous thing: What the public thinks and knows about paying for college*. Washington, DC: American Council on Education.

Kane, Thomas J. 1995. Rising public college tuition and college entry: How well do public subsidies promote access to college? NBER Working Paper No. 5164.

Kane, Thomas J. 2003. A quasi-experimental estimate of the impact of financial aid on college-going. NBER Working Paper No. 9703.

Kane, Thomas J. 2004. Evaluating the impact of the D.C. Tuition Assistance Grant Program. NBER Working Paper No. 10658.

Kane, Thomas J. 2007. Evaluating the impact of the D.C. Tuition Assistance Grant Program. *Journal of Human Resources* 42(3): 555–82.

Kinsley, Peter, and Sara Goldrick-Rab. 2011. How financial constraints affect college enrollment intensity decisions. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April.

Knetsch, Jack L., and J. A. Sinden. 1984. Willingness to pay and compensation demanded: Experimental evidence of an unexpected disparity in measures of value. *Quarterly Journal of Economics* 99(3): 507–21.

Lawrence, Emily. 1991. Poverty and the rate of time preference. *Journal of Political Economy* 99(1): 54–77.

Leslie, Larry, and Paul Brinkman. 1988. *The economic value of higher education*. New York: Macmillan.

Linsenmeier, David M., Harvey S. Rosen, and Cecilia Elena Rouse. 2006. Financial aid packages and college enrollment decisions: An econometric case study. *Review of Economics and Statistics* 88(1): 126–45.

Manski, Charles, and David A. Wise. 1983. *College choice in America*. Cambridge, MA: Harvard University Press.

Mathiowetz, Nancy A., Charlie Brown, and John Bound. 2002. Measurement error in surveys of the low-income population. In *Studies of welfare populations: Data collection and research issues*, edited by Michele Ver Ploeg, Robert A. Moffitt, and Constance F. Citro, pp. 157–94. Washington, DC: National Academies Press.

Odean, Terrance. 1998. Are investors reluctant to realize their losses? *Journal of Finance* 53(5): 1775–98.

Paulsen, Michael B., and Edward P. St. John. 1997. The financial nexus between college choice and persistence. In *Researching student aid: Creating an action agenda*, edited by R. A. Voorhees, pp. 65–82. San Francisco, CA: Jossey-Bass.

Paulsen, Michael B., and Edward P. St. John. 2002. Social class and college costs: Examining the financial nexus between college choice and persistence. *Journal of Higher Education* 73(3): 189–236.

Perna, Laura. 2006. Understanding the relationship between information about college prices and financial aid and students' college-related behaviors. *American Behavioral Scientist* 49(12): 1620–35.

Raudenbush, Stephen W. 2011. Modeling mediation: Causes, markers, and mechanisms. Paper presented at the Annual Society for Research on Educational Effectiveness Conference, Washington, DC, March.

Reyes, Suzanne L. 1995. Educational opportunities and outcomes: The role of the guaranteed student loan. Unpublished paper, Harvard University.

Schneider, Barbara, and Sarah-Kathryn McDonald. 2006. *Scale up in education: Issues in practice*. New York: Rowan and Littlefield.

Schrag, Francis. 1989. Values in educational inquiry. *American Journal of Education* 97(2): 171–83.

Scott-Clayton, Judith. 2011. On money and motivation: A quasi-experimental analysis of financial incentives for college achievement. *Journal of Human Resources* 46(3): 614–46.

Scrivener, Susan, and Jenny Au. 2007. Enhancing student services at Lorain County Community College: Early results from the Opening Doors Demonstration in Ohio. New York: MDRC.

Scrivener, Susan, and Michael Pih. 2007. Enhancing student services at Owens Community College: Early results from the Opening Doors Demonstration in Ohio. New York: MDRC.

Seftor, Neil, and Sarah Turner. 2002. Back to school: Federal student aid policy and adult college enrollment. *Journal of Human Resources* 37(2): 336–52.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Shefrin, Hersch, and Meir Statman. 1985. The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance* 40(3): 777–90.

Small, Mario L. 2009. Lost in translation: How not to make qualitative research more scientific. In *Workshop on interdisciplinary standards for systematic qualitative research*, edited by Michèle Lamont and Patricia White, pp. 165–71. Washington, DC: National Science Foundation.

Stanley, M. 2002. College education and the mid-century G.I. Bills. *Quarterly Journal of Economics* 118(2): 671–708.

Thaler, Richard. 1981. Some empirical evidence on dynamic inconsistency. *Economics Letters* 81(3): 201–7.

Thaler, R. and Sunstein, C. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. New Haven: Yale University Press.

Turner, Sarah, and John Bound. 2003. Closing the gap or widening the divide: The effects of the G.I. Bill and World War II on the educational outcomes of black Americans. *Journal of Economic History* 63(1): 145–77.

van der Klaauw, Wilbert. 2002. Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review* 43(4): 1249–88.

Yoshikawa, Hirokazu, Thomas S. Weisner, Ariel Kalil, and Niobe Way. 2008. Mixing qualitative and quantitative research in developmental science: Uses and methodological choices. *Developmental Psychology* 44(2): 344–54.

Yoshikawa, Hirokazu, Thomas S. Weisner, and Edward D. Lowe. 2009. *Making it work: Low-wage employment, family life, and child development*. New York: Russell Sage Foundation.

Zelizer, Viviana. 1994. *The social meaning of money*. New York: Basic Books.