Journal of Postsecondary Student Success 3(1)
doi: 10.33009/fsop_jpss132759

Empirical Benchmarks for Planning and Interpreting Causal Effects of Community College Interventions

Michael J. Weiss[*], Marie-Andrée Somers, & Colin Hill

MDRC

Abstract

Randomized controlled trials (RCTs) are an increasingly common research design for evaluating the effectiveness of community college (CC) interventions. However, when planning an RCT evaluation of a CC intervention, there is limited empirical information about what sized effects an intervention might reasonably achieve, which can lead to under- or over-powered studies. Relatedly, when interpreting results from an evaluation of a CC intervention, there is limited empirical information to contextualize the magnitude of an effect estimate relative to what sized effects have been observed in past evaluations. We provide empirical benchmarks to help with the planning and interpretation of community college evaluations. To do so, we present findings across well-executed RCTs of 39 CC interventions that are part of a unique dataset known as The Higher Education Randomized Controlled Trials (THE-RCT). The analyses include 21,163–65,604 students (depending on outcome and semester) enrolled in 44 institutions. Outcomes include enrollment, credits earned, and credential attainment. Effect size distributions are presented by outcome and semester. For example, across the interventions examined, the mean effect on cumulative credits earned after three semesters is 1.14 credits. Effects around 0.16 credits are at the 25th percentile of the distribution. Effects around 1.69 credits are at the 75th percentile of the distribution. This work begins to provide empirical benchmarks for planning and interpreting effects of CC evaluations. A public database with effect sizes is available to researchers (https://www.mdrc.org/the-rct-empirical-benchmarks).

Keywords: empirical benchmarks, effect size, community college, postsecondary, enrollment, credits, degree attainment, randomized controlled trial

[*] Contact: Michael.Weiss@mdrc.org

Empirical Benchmarks for Planning and Interpreting Causal Effects of Community College Interventions

Imagine you are a reviewer of a proposal seeking to evaluate the effectiveness of a comprehensive intervention to support community college (CC) students. The research team proposes a 600-student randomized controlled trial (RCT) evaluation of the intervention. The proposal includes a presentation of the minimum detectable effect, that is, "the smallest true impact that an experiment has a good chance of detecting" (Bloom, 1995, p. 547). With 600 students, the minimum detectable effect on college degree completion is 8 percentage points. Is this the right sized effect to target? Might the intervention realistically have an effect of this magnitude?

Or, imagine that you are a college administrator reading about the findings from an evaluation of the Detroit Promise Path (DPP) intervention. DPP provides graduates of Detroit high schools attending a local CC with a dedicated coach and $50 each month if they meet with their coach twice per month. To find out if DPP is effective, researchers conducted an RCT evaluation. One year after students joined the evaluation, those offered DPP earned 8.9 credits, on average. Those in the control group, who were not offered DPP, earned 7.1 credits, on average. The estimated average effect, or value-added, of DPP was thus 1.8 credits earned ($p = 0.0007$; Ratledge et al., 2021). How do we interpret this finding? Is an effect of 1.8 credits accumulated through one year small, medium, or large?

The present paper[1] presents one type of empirical benchmark that can be used for both planning evaluations of CC interventions, like the comprehensive support program above, and interpreting causal effect estimates from evaluations of CC interventions, like DPP. Conceptually, empirical benchmarks can be helpful to frame the magnitude of an intervention's effects relative to something else. In educational evaluations, examples of such comparative benchmarks include: (a) the distribution of estimated effects from evaluations of other related interventions; (b) normative expectations for educational progress (e.g., typical student growth on achievement tests during the year); (c) prevailing gaps in educational outcomes (e.g., racial inequality in academic achievement); (d) policy-relevant performance thresholds (e.g., the probability of being proficient on a state test); and (e) cost-effectiveness ratios (for examples, see Baird & Pane, 2019; Hill et al., 2008; Kraft, 2020). The present paper focuses on the first type of empirical benchmark.

In doing so, we aim to support researchers planning an evaluation, and funders who might support them, to ensure that evaluations of CC interventions are adequately powered to detect effects that might realistically be achieved, given what has been observed in the past. We also aim to support researchers and policymakers seeking

[1] Portions of this text, such as descriptions of the studies and data sources, are adapted from Somers et al. (2023) and Weiss et al. (2022).

to interpret an estimated effect of a CC intervention, by providing one valuable piece of context—how effective their intervention is relative to the effectiveness of other rigorously evaluated CC interventions.

Background

CCs play a vital role in U.S. postsecondary education. In fall 2021, 4.5 million students attended public two-year colleges, representing 29% of U.S. undergraduates.[2] Despite providing unprecedented access to postsecondary education, rates of degree attainment remain low. Only 31% of first-time, full-time students seeking a degree or certificate whose first postsecondary school is a public two-year college graduate within three years (Integrated Postsecondary Education Data System (IPEDS) Trend Generator, 2021). To address these issues, policymakers, foundations, and college administrators are beginning to embrace the need for causal evidence of the effectiveness of postsecondary programs, policies, and practices.

In 2002, the U.S. Department of Education created the Institute for Educational Sciences (IES) as part of the Education Sciences Reform Act, which has provided unprecedented funding for educational evaluations with strong potential to draw causal conclusions. Thus began a transformation in higher education evaluation. Two decades later, MDRC alone has conducted 31 RCTs of 41 interventions in over 45 (mostly community) colleges throughout the United States, including 67,400 students, mostly from low-income backgrounds (Diamond et al., 2021). Many more RCTs in higher education have been conducted by others. For example, the What Works Clearinghouse (WWC) has published reviews of 68 large-scale (with more than 350 participants) postsecondary RCTs that meet their evidence standards without reservations.[3]

While the number of RCTs in CCs has grown dramatically over the past 20 years, the information needed to plan a high-quality RCT and to interpret their findings in this context has not. When planning an RCT, it is important to consider the size of effect that the intervention might reasonably achieve. Researchers can use this information when setting a target sample size needed to ensure a study is adequately powered. Relatedly, when interpreting RCT findings, it is important for researchers to convey the practical significance of effect estimates to policymakers and practitioners to help inform their decision-making. In both scenarios, researchers need information that

[2] Source: U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System (IPEDS), Fall Enrollment component final data (2002–2020) and provisional data (2021).

[3] The 68 postsecondary RCTs were found through https://ies.ed.gov/ncee/wwc/StudyFindings (accessed 09/24/22) under topic area "Postsecondary," excluding "secondary mathematics" and "college transition" programs, where the study sample was postsecondary.

will help them consider what effect sizes are meaningful and policy relevant in the context of their study.

For many years, evaluators would plan RCTs and interpret their findings using the benchmarks proposed (as somewhat of a last resort) by Cohen for "small" (0.2), "medium" (0.5), and "large" (0.8) effect sizes. However, more recently in K-12 research, one powerful approach that has been adopted to characterize the magnitude of an effect of an educational intervention is to compare it with previously estimated effects of interventions in a similar context (Hill et al., 2008). Based on an analysis of prior K-12 studies, Hill et al. (2008) found that the average estimated effect size was 0.07, with a 0.32 standard deviation. Assuming approximate normality of the distribution of estimated effects, these findings indicate that, in the elementary school context on broad standardized tests, only around 1% of the examined estimated effect sizes were large by Cohen's rules of thumb (suggesting those rules of thumb are probably not appropriate in the K-12 context). The empirical findings of Hill et al. (2008) have changed the way K-12 researchers plan studies (expected effect sizes have decreased) and interpret their findings (what used to be thought of as "small" effects are now taken quite seriously).

Based on Hill et al.'s (2008) work and numerous studies since, K-12 researchers now have access to empirical benchmarks for standardized effect size estimates based on results from real-world RCTs (Baird & Pane, 2019; Bloom et al., 2008; Hill et al., 2008; Kraft, 2020, 2023; Lipsey et al., 2012; Wolf & Harbatkin, 2022). These benchmarks can be used to situate an effect estimate within the distribution of effect estimates observed in K-12 evaluations, by subject area and grade-level, and even by other factors affecting the magnitude of intervention effects such as outcome type and study design. Although these normative benchmarks do not necessarily inform what effects are practically meaningful to decision-makers (for this purpose, Hill et al., 2008 and others have proposed other benchmarks, discussed in the conclusion), they are very useful for grounding expectations for what might realistically be attainable when conducting power calculations or interpreting the magnitude of a study's effects (Konstantopoulos & Hedges, 2008, p. 1615).

In stark contrast, in postsecondary education, such information is not available. This may partially explain why many postsecondary RCTs appear to be underpowered.[4] Thus, the purpose of the present paper is to provide normative empirical benchmarks for postsecondary interventions, like the types that already exist in K-12 education. Our focus is on the intervention-level distribution of *average effects* (mean, standard deviation, and percentiles) across 39 postsecondary (mostly CC) interventions evaluated using an RCT. We examine the distribution of effects on multiple outcomes typically examined in CC studies (enrollment, credit accumulation, degree completion)

[4] Reviews have found that many education studies (and not just CC RCTs) are not adequately powered, see Cheung & Slavin (2016), Somers et al. (2023), Spybrook et al. (2020), and Torgerson et al. (2005).

and by semester (from one through six semesters after individuals joined each study). The distributions presented can be used for planning future evaluations, by helping researchers determine what sample size will be needed to detect effects that might plausibly be achieved, given the intervention they intend to evaluate and the outcomes they plan to measure. They can also help inform the interpretation of impact findings from evaluations like the DPP example introduced earlier.

In addition to providing these benchmarks in a new context, our methodological approach diverges from past efforts in a way that we believe represents an important improvement. Historically, researchers have presented the distribution of *estimated* effects from a group of studies for benchmarking. However, due to "estimation error variance," the distribution of *estimated* effects is known to vary more, and sometimes a lot more, than the distribution of *true* effects (Bloom et al., 2017, p. 824). The approach we use aims to remove estimation error variance from the distribution of estimated effects, allowing for estimation of how much *true* effects vary among interventions, thus producing more appropriate benchmarks for planning and interpretation purposes.

With respect to planning, and more specifically power calculations, the "minimum detectable effect (MDE)" would more accurately be called the "minimum detectable *true* effect (MDTE)"—it is not the minimum detectable *estimated* effect (Somers et al., 2023). Consequently, when planning a study and considering whether the MDTE is of a magnitude that might be achieved by the intervention under study, a relevant benchmark is not "where does the MDTE lie on the distribution of *estimated* effects from past studies?" Instead, it is "where does the MDTE lie on the estimated distribution of *true* effects from past studies?" Our proposed benchmarks more accurately support answering this question.

For interpretation purposes, what again seems most relevant is our best understanding of where the effect of an intervention under consideration is situated within the distribution of *true* effects from past evaluations, not within the wider distribution of effect *estimates* from past evaluations.[5] Our proposed approach should mitigate some of the challenges associated with "promising trials bias" and "the winner's curse" (Simpson, 2022; Sims et al., 2022). We illustrate how the distinction between the distribution of *estimated* effects and the estimated distribution of *true* effects can affect study planning and interpretation, by comparing the two approaches on one of our outcomes of interest.

In the next section we describe the methodology used in our analysis, including data sources, measures, estimands, and statistical models. In the following section we share results. In the last section we offer a discussion of key implications and limitations of those findings.

[5] All else equal, the smaller the size of the underlying studies used to develop the empirical benchmarks, the more important the distinction between estimated effects and true effects becomes. The present empirical benchmarks are derived from fairly large-scale evaluations (n > 700 in 38 out of 39 evaluations). Thus, the significance of the methodological improvements in this research may be smaller in magnitude than had more of the underlying studies been smaller.

Methods

Studies and Analysis Samples

Our analyses focus on evaluations of CC interventions where the identification strategy allows for an unbiased estimator of intervention effects. In doing so, we ensure that the distribution of effects is not confounded with cross-study variation in degrees of bias. In addition, we aimed to identify evaluations that are representative of the CC interventions that have been rigorously evaluated to date, to optimize the comprehensiveness and generalizability of the benchmarks.

Accordingly, the findings in this paper are based on 30 well-executed RCTs of postsecondary interventions conducted by MDRC, which represent all but one of the postsecondary RCTs that MDRC had led from 2003–2019 (the one RCT excluded from the present analysis had limited follow-up). We present findings on student outcomes for the first six semesters after random assignment, a common time to consider CC degree completion rates, thereby making it possible to examine the pattern of variation in effect sizes across semesters of follow-up as well as across interventions.

Importantly, the impact findings from these RCTs are causally robust. Twenty-seven of these RCTs have been reviewed by the U.S. Department of Education's WWC and have all met the WWC's evidence standards without reservations. The three RCTs that have not yet been reviewed almost certainly meet the same standards, given their similar design, analytic approach, and attrition rates.

Furthermore, the RCTs used for this paper comprise a sizable portion of all large-scale postsecondary RCTs conducted in the United States. Only 68 large-scale (with more than 350 participants) postsecondary RCTs have been reviewed and met WWC evidence standards without reservations. Thus, the findings from this paper likely provide a representative picture of the distribution of effect sizes in well-executed CC RCTs.[6]

Some of the 30 RCTs used in this paper are multi-arm trials. These multi-arm trials evaluate the effect of more than one intervention in a single RCT by, for example, randomly assigning students to a control group, intervention A, or intervention B. Thus, although there were 30 RCTs in our sample, they are used to estimate the distribution of effects of 39 interventions.[7] The resulting full study sample includes 39 postsecondary interventions and a total of 65,604 students.

[6] See footnote 3.

[7] In the multi-arm trials, for analysis purposes students in the common control group were randomly divided (within blocks) into as many groups as there were intervention arms, thus creating a unique control group for each intervention arm in the RCT. This avoids having to deal with analytic complications arising from having a shared control group. This approach follows Weiss et al. (2022).

As shown in Table 1, the 39 studied interventions vary in their key components (e.g., advising, tutoring, financial supports, etc.) and duration (from one semester to three years). For more information on the key components of each individual intervention, see Appendix Tables A2 and A3. For even greater detail, Appendix Table A1 provides links to original reports about each intervention.

The eligible population in most studied interventions was students enrolled in the colleges (as opposed to prospective students or applicants); in two thirds of interventions, eligibility was limited to new or first-year students (see Table 2). Most interventions were implemented at CCs or public universities with large populations of students from families with low-income (or the interventions targeted students from families with low-income) and most studies included multiple cohorts. The interventions were implemented across 44 postsecondary institutions (mostly CCs) and 12 states.

Table 1. Intervention Components and Duration

Intervention Characteristic	Percentage of Interventions
Presence of component	
Enhanced advising	38%
Enhanced tutoring	28%
Financial support	51%
Instructional reform	26%
Learning communities	23%
Promoting full-time/summer enrollment	33%
Success course	23%
Duration (Years)	
0.5 Year	38%
1.0 Year	36%
1.5 Years	8%
2.0 Years	8%
2.5 Years	0%
3.0 Years	8%
Number of interventions:	39

Note. One intervention was financial aid reform that did not result in any increase in the amount of aid distributed. It is therefore the only intervention with none of the seven intervention components that were coded. Sources: MDRC calculations using data from THE-RCT and reports and journal articles. A list of reports and articles can be found in Appendix Table A1.

Table 2. Eligibility Criteria and Study Sample Sizes

Study Characteristics	Percentage of Interventions
Student eligibility criteria [a]	
Low-income	54%
Remedial needs	36%
New or first year	59%
Enroll full time	10%
Other	82%
Study sample size	
700 or fewer	3%
701–1,000	31%
1,001–2,000	21%
2,001–5,000	44%
5,000 or greater	3%
Number of interventions:	39

Note. Interventions are equally weighted. Sources: MDRC calculations using data from THE-RCT and reports and journal articles. A list of reports and articles can be found in Appendix Table A1.

[a] Percentages do not add up to 100% because interventions may have more than one eligibility criteria.

Reflecting national patterns in two-year colleges, most students (77%) in the average study are younger than 25 (see Table 3). Almost two thirds of students (60%) in the average study are female, and the average percent Black is 25%, the average percent Hispanic is 36%; both percentages are higher than in the average two-year college in the United States. There is substantial variation in the characteristics of students across the studies; for example, the percentage of female students ranges from 0%–92%, and the percentage of White students ranges from 0%–60%.

Table 3. Characteristics of Students in the Average Study in the Main Analytic Sample

Student Characteristics	Percent (mean) across interventions	Range across Interventions
Gender		
Female	60%	0%–92%
Male	39%	8%–100%
Missing	1%	0%–15%

Table 3. Characteristics of Students in the Average Study in the Main Analytic Sample (*continued*)

Student Characteristics	Percent (mean) across interventions	Range across Interventions
Racial-ethnic group		
Black	25%	0%–82%
Hispanic	36%	2%–100%
White	26%	0%–60%
Asian	5%	0%–13%
Other	4%	0%–13%
Missing	4%	0%–18%
Age		
Younger than 25	77%	30%–100%
25 or older	22%	0%–70%
Missing	0%	0%–2%
Number of interventions	39	39

Note. Interventions are equally weighted. Sources: MDRC calculations using data from THE-RCT and reports and journal articles. A list of reports and articles can be found in Appendix Table A1.

The analytic sample used to estimate the distribution of effects varies across outcomes and by semesters depending on data availability, ranging from 20 to 39 interventions, and 21,163 to 65,604 students.[8] Table 4 shows the percentage of individuals and interventions that are included in the analysis of effects for each outcome and semester, relative to the full study sample. In semesters 1–3, at least 82% of studied interventions and students are included in the analysis; however, in semesters 4–6, the number of studied interventions with longer-term follow-up data drops. For example, about two thirds of studies (26 studied interventions) collected enrollment data in semester 6 and half of studies collected credits and degree completion data (22 and 20 studied interventions, respectively) through semester 6. For this reason, caution is needed when interpreting results for outcomes beyond semester 3. The distribution of effects is very likely upward biased due to follow-up selection bias. That is, interventions with more promising short-term impacts were more likely to have longer-term follow-up data (see Bailey & Weiss, 2022).

[8] Although there are 65,637 students in the data, the largest analytic sample is 65,604 because 33 students in the PBS NY study are missing some outcome data due to a matching issue in the original data collection.

Table 4. Data Availability as a Percent of the Full Sample of Individuals and Interventions

Outcome Measure and Unit	Semester					
	1	2	3	4	5	6
Enrollment						
Individuals	100%	100%	88%	83%	66%	51%
Interventions	100%	100%	97%	97%	79%	67%
Credits Earned						
Individuals	92%	92%	85%	68%	53%	32%
Interventions	85%	85%	82%	77%	74%	56%
Degrees Earned						
Individuals				61%	56%	47%
Interventions				59%	59%	51%

Note. The full sample includes 65,637 students and studies of 39 interventions. Sample sizes by outcome and semester are shown in Appendix Table A4.

Because the data used for the present analysis are from actual RCTs—nearly all of which focus on students who are already enrolled and who agreed to participate in the study—the distribution of effects presented in this paper may not generalize to the effects that one would observe if these interventions were offered to all CC students at the colleges (or in the United States), nor to prospective students or applicants to these colleges. However, the findings are likely to represent the range of effect sizes that researchers will encounter for the subset of colleges and enrolled students who are interested in these types of interventions.

Data Sources and Measures

Data Sources

The data for this paper are from MDRC's *The Higher Education Randomized Controlled Trials Restricted Access File* (*THE-RCT RAF*; Diamond et al., 2021). THE-RCT RAF is a restricted access student-level database created by MDRC and housed at the University of Michigan's Inter-university Consortium for Political and Social Research (ICPSR). The database includes all RCTs that MDRC has conducted in postsecondary education from 2003–2019 and is available to qualified researchers with few restrictions.

The data includes information about each RCT's design (e.g., study name, experimental group indicators, and random assignment block indicators) plus students' characteristics and their academic outcomes by semester (enrollment, credit accumulation, degree completion). These data were originally obtained from three sources: (a) college (or college system) records, which include demographic records, course transcripts, and degree completion; (b) the National Student Clearinghouse (NSC), which maintains information on enrollment and degree completion from nearly 3,600 colleges that combined enroll over 97% of the nation's college students (https://www.studentclearinghouse.org/about/); and (c) study-administered student surveys implemented at the time of random assignment, to collect information on student characteristics that are not available from college records.[9]

Outcome Measures

The outcomes explored in the present analysis are the focus of most CC interventions:

- *Persistence (enrollment).* To make academic progress, students must continue to enroll in college over time. We examine the percentage of students who were enrolled in postsecondary education each semester, as well as the cumulative number of semesters enrolled by a given semester.[10]
- *Total credits accumulated.* Total credits accumulated is a critical indicator of students' academic progress toward a degree, which typically requires at least 60 college-level credits (in the CC setting). Consequently, we examine total credits earned by semester and cumulatively during students' first six semesters after random assignment.[11]

[9] The present paper uses Version 1.0 of the THE-RCT RAF database, which we supplemented in three ways for the purposes of our analysis. First, Version 1.0 of THE-RCT RAF excludes individual-level data from five of the RCTs included in the present paper (all the studies conducted at the City University of New York); these student-level data were included in the present analysis and are included in Version 4.0 of THE-RCT RAF. Second, Version 1.0 of the RAF included a shorter follow-up for the Detroit Promise Path RCT, EASE, and ModMath—for the present analysis we were able to include updated data through six semesters of follow-up for the full samples. Finally, the "credits earned" variable for one study (CUNY Start) is not available in THE-RCT RAF. Its derivation is described in Weiss et al. (2022), Appendix B.

[10] We emphasize enrollment in the second through sixth semesters after random assignment because few of the interventions intended to impact first semester enrollment—these interventions largely targeted students who already intended to enroll in semester 1 (e.g., new or continuing students who had already registered). Notable exceptions include: the Encouraging Additional Summer Enrollment (EASE) interventions, Detroit Promise Path (DPP), and the Performance-Based Scholarships (PBS) Variations interventions.

[11] Most community colleges in THE-RCT offered a longer "main" session in the fall and spring and a shorter "intersession" in the winter and summer. For the purposes of this analysis, we group the fall and winter sessions into a "semester" and the spring and summer sessions into another "semester," so that each year of post random assignment follow-up is comprised of two semesters. To maximize the length of follow-up, if the final session of data available for a student is a fall or spring session, we treat that as the data for the full semester. For example, if data are available for a student for three years except the final winter or summer intersession, we still treat this as a semester six outcome.

- *Degree or certificate completion.* Our measure of this marker is the percentage of students who earned a postsecondary credential by four to six semesters after random assignment, common time frames for measuring CC degree completion.

Data on the number of credits earned are from college or system records. Information on enrollment and degree completion are from college or system records data or from the NSC, depending on the study and the student. To maximize data availability across studies, these outcomes are derived using all data sources. When only college/system outcome data are available for a study, these measures are defined as enrollment or degree completion at the college/system of random assignment. When both college/system and NSC data are available for a study, these outcomes are defined as enrollment or degree completion at any college/university covered by the two sources. This means that enrollment and degree completion are measured somewhat differently across studies (either at nearly any college in the nation or only the college/system where the study took place, depending on data availability for the study).[12]

Attrition of sample members does not present a problem in our analyses. For enrollment, credit accumulation, and degree completion, data are available for nearly every student in the study, if the relevant information (e.g., transcript records for credit accumulation) was collected for that study.[13] When the college or NSC data include no records for a given student, we treat that student as not being enrolled, and therefore earning zero credits and not earning a degree. (The sample size reductions shown in Table 3 across semesters are due to a shorter follow-up period for some studies or cohorts, rather than sample attrition for other reasons.)

Supplemental findings in a database associated with this paper include an examination of the distribution of effects on more narrowly defined versions of the main outcomes, including full-time enrollment, developmental credits earned, college-level credits earned, as well credits attempted (total, developmental, and college-level). As an additional supplemental analysis, we also examine effects on students' performance in their courses as measured by their grade point average (GPA), measured on a 4-point scale. Information on GPA is from college records. Impacts on GPA are challenging to evaluate in postsecondary impact studies because GPA is only defined for students who are still enrolled. This means that if an intervention has an impact on enrollment, the

[12] As a sensitivity check we estimated each intervention's impact on earning a degree at the college/system of random assignment versus at any college or university in NSC for those studies where both data sources were available. There are only very small differences between effect estimates using these different data sources (less than ½ a percentage point). Thus, this does not appear to be a major concern.

[13] After being randomly assigned, some students asked to be removed from the study. Rates of overall attrition were 2.8% and below in all studies except PBS Variations, which had an overall rate of attrition of 4.6%.

estimator of the effects on GPA could be biased (and in all cases, estimated effects on GPA do not apply to unenrolled students). For the present analysis, comparing impacts on GPA across follow-up semesters is also challenging because non-enrollment from the sample increases over time. Therefore, the findings for GPA in the online database are limited to the first follow-up semesters.

Parameters

Before delving into how we estimate key parameters of interest from our data, we first define two key parameters of interest. Let B_j be the *true* average effect of intervention j. Our analyses begin by estimating two parameters that summarize the cross-intervention distribution of intervention mean effects—its mean (β) and its standard deviation (τ). By definition:

$$\beta \equiv \lim_{J^* \to \infty} \frac{\sum_{j=1}^{J^*} B_j}{J^*} \tag{1}$$

and

$$\tau \equiv \sqrt{Var\left(B_j\right)} \equiv \lim_{J^* \to \infty} \sqrt{\frac{\sum_{j=1}^{J^*} (B_j - \beta)^2}{J^*}}. \tag{2}$$

β and τ provide a summary of the central tendency and spread of the distribution of average *true* effects across interventions, respectively.

In addition to these two primary summary statistics, we aim to characterize the distribution of *true* effects by identifying points on the distribution that correspond with percentiles of the distribution.

Important Context about Distributions of Estimated Effects

Much of the K-12 education (and other fields) literature base on empirical benchmarks starts with a group of studies and the *estimated* effects of the interventions from those studies. Empirical benchmarks often include the mean, median, standard deviation, and various percentiles of the distribution of the *estimated* effects from those studies (for examples, see Bloom et al., 2008; Hill et al., 2008; Kraft, 2020; Lipsey et al., 2012). Notice the emphasis on *estimated*.

Define $\hat{B}_j^{Orig}$ as an original *estimate* of the average effect of intervention j, estimated using an unbiased estimator (e.g., a simple difference-in-means estimator or a regression-based estimator). Such estimates ($\hat{B}_j^{Orig}$) are a combination of the *true* effect (B_j) and random estimation error (r_j). That is:

$$\hat{B}_j^{Orig} = B_j + r_j \,. \tag{3}$$

Consequently, the spread (or variance) of the distribution of effect *estimates* is a combination of variation in the *true* effects of the interventions ($Var\left(B_j\right)$) plus independent variation in estimation error ($Var\left(r_j\right)$). Accordingly, the distribution of effect *estimates* is expected to be wider (and sometimes much wider) than the distribution of *true* effects (for more details, see Bloom et al., 2017 and Hedges & Pigott, 2001). That is:

$$Var\left(\hat{B}_j^{Orig}\right) = Var\left(B_j\right) + Var\left(r_j\right), \tag{4}$$

which can be re-written as:

$$Var\left(\hat{B}_j^{Orig}\right) = \tau^2 + Var\left(r_j\right). \tag{5}$$

Recall from Equation 2 that τ, the standard deviation of the cross-intervention distribution of true average effects, is a key parameter of interest. The standard deviation of $\hat{B}_j^{Orig}$, as is commonly presented in the literature, overestimates the spread of the distribution of *true* effects. Relatedly, percentiles of the distribution of estimated effects present an inaccurate depiction of percentiles on the distribution of true effects. We attempt to address this issue with our chosen estimators.

Estimators

To examine the distribution of *true* effects across studied interventions, we use the fixed-intercept, random treatment coefficient (FIRC) model described in detail by Bloom et al. (2017) for studying cross-site impact variation.[14] The FIRC approach was used by Weiss et al. (2017) for their secondary analysis of data from 16 multi-site RCTs of education and training programs.

Specifically, we use the following 2-level hierarchical linear model to estimate β and τ:

Level 1: Sample Members

$$Y_{ij} = \boldsymbol{\acute{a}} \cdot \boldsymbol{S}_{ij} + B_j T_{ij} + e_{ij}\ , \tag{6}$$

Level 2: Studied Interventions

$$B_j = \beta + b_j, \tag{7}$$

where:

$$e_{ij} \sim N\left(0,\left(1-T_{ij}\right)\sigma^2_{0|S} + T_{ij}\sigma^2_{1|S}\right)$$

[14] Bloom et al. (2017) received the 2017 Outstanding Article Award from the *Journal of Research on Educational Effectiveness.*

$$b_j \sim N\left(0,\tau^2\right)$$

$$Cov\left(e_{ij},b_j\right)=0.$$

In this model, Y_{ij} is the value of the outcome (e.g., credits earned) for individual *i* in studied intervention *j*, $\boldsymbol{S}_{ij}$ is a vector of random assignment block indicators (one for each block in each studied intervention) set equal to one if student *i* in studied intervention *j* was randomly assigned in that random assignment block and zero otherwise, T_{ij} equals one if individual *i* in studied intervention *j* was assigned to treatment and zero otherwise. The blocks account for the fact that individuals were randomly assigned within blocks (e.g., colleges and cohorts) and that the proportion of sample members randomized to treatment can vary across blocks. The model does not control for students' baseline characteristics (like their gender and age). Doing so would not appreciably improve the precision of estimated effects because available characteristics are only weakly correlated with outcomes (Somers et al., 2023), and very few baseline variables are consistently available across studies.

An important feature of the FIRC model, relevant to the purpose of this paper, is that it allows for intervention-specific effect coefficients (B_j) that can vary randomly across studied interventions. The B_j's are modeled as representing a cross-intervention population distribution with a mean value of β and a standard deviation of τ. Hence, the intervention-level random error term, b_j, has a mean of zero and a standard deviation of τ. Critically, when estimating τ, FIRC accounts for estimation error associated with each intervention specific effect estimate.[15] Bloom et al. (2017) provide further information about this model and Raudenbush and Bloom (2015) explore its properties.

For the present analysis, the FIRC model is fitted to the analysis samples separately for each outcome and semester. Estimates of β (average) and τ (standard deviation) are key summaries of findings.

To further aid with interpretation, we also provide percentiles of the distribution of intervention effects. As previously noted, because of estimation error a key challenge with calculating percentiles is that the distribution of estimated study-specific effects (e.g., the $\hat{B}_j^{Orig}$ reported in the original studies) exaggerates the amount of *true* cross-study variation in effects (Bloom et al., 2017; Raudenbush & Bloom, 2015). We address this problem using a two-pronged approach proposed by Bloom et al. (2017). First, we begin by using the results of the FIRC model to compute the empirical Bayes shrinkage impact estimate for each intervention, $\hat{\beta}_j^{EB}$, which is a weighted average of

[15] The model also allows for the variability of level-1 residuals to differ by treatment group. The individual-level random error term, e_{ij}, is assumed to have a mean of zero and a variance of $\left(1-T_{ij}\right)\sigma^2_{0|S}+T_{ij}\sigma^2_{1|S}$, which can be different for treatment group members and control group members.

the intervention-specific average impact estimate, $\hat{B}_j^{Orig}$, and the overall average impact estimate, $\hat{\beta}$, where the weight of the study-specific estimate is based on its reliability, λ_j:[16]

$$\hat{\beta}_j^{EB} = \lambda_j \hat{B}_j^{Orig} + \left(1 - \lambda_j\right)\hat{\beta} \tag{8}$$

This means that for small-sample studies, where estimated effects are estimated less reliably, the empirical Bayes estimates will be "shrunken" towards the grand mean impact estimate. The shrinkage factor is based on an estimate of reliability. The resulting distribution of empirical Bayes effect estimates varies less than the best estimate of the variance of the distribution of true effects (Raudenbush & Bryk, 2002, p. 88). That is, $Var\left(\hat{\beta}_j^{EB}\right) < \hat{\tau}^2$. Thus, the variance of empirical Bayes estimates will typically understate the cross-study variance of true mean program effects, and by extension, be smaller than the estimated cross-study variation from the FIRC model ($\hat{\tau}^2$). Hence, as a second step, we calculate an "adjusted" empirical Bayes estimate for each study to compensate for this over-shrinkage (Bloom et al., 2017):

$$\hat{\beta}_j^{AEB} = \hat{\beta} + \frac{1}{\sqrt{\gamma}}\left(\hat{\beta}_j^{EB} - \hat{\beta}\right), \tag{9}$$

where γ is an adjustment factor that stretches the distance between the empirical Bayes estimates and the mean impact estimate $\hat{\beta}$:

$$\gamma = \frac{\widehat{var}(\hat{\beta}_j^{EB})}{\hat{\tau}^2}.$$

This adjustment inflates the variance of the empirical Bayes estimates to be exactly equal to the estimated variance of true program effects ($\hat{\tau}^2$) from the FIRC model. The percentiles presented in this paper are based on the adjusted empirical Bayes estimates, $\hat{\beta}_j^{AEB}$. Study-specific estimates are also available in a public-use dataset created for this paper (available at https://www.mdrc.org/the-rct-empirical-benchmarks).

Results

Distribution of Impact Estimates—An Example Using Three Estimators

Before presenting our main results, it is useful to examine the distribution of intervention impact estimates for a single outcome at a single time point, highlighting some important points about our methodological approach. Figure 1 presents the estimated impact of each intervention on the cumulative number of semesters enrolled through two semesters after random assignment. Impact estimates are presented using three

[16] See Bloom et al. (2017) for details on calculating λ_j.

estimators: (a) original impact estimates (called $\hat{B}_j^{Orig}$, above, and "OLS" in Figure 1)[17], (b) empirical Bayes impact estimates (called $\hat{B}_j^{EB}$, above), and (c) adjusted empirical Bayes impact estimates (called $\hat{B}_j^{AEB}$, above). Interventions are listed in the order of the magnitude of their adjusted empirical Bayes impact estimate. The horizontal axis of the figure indicates the direction and magnitude of each impact estimate.

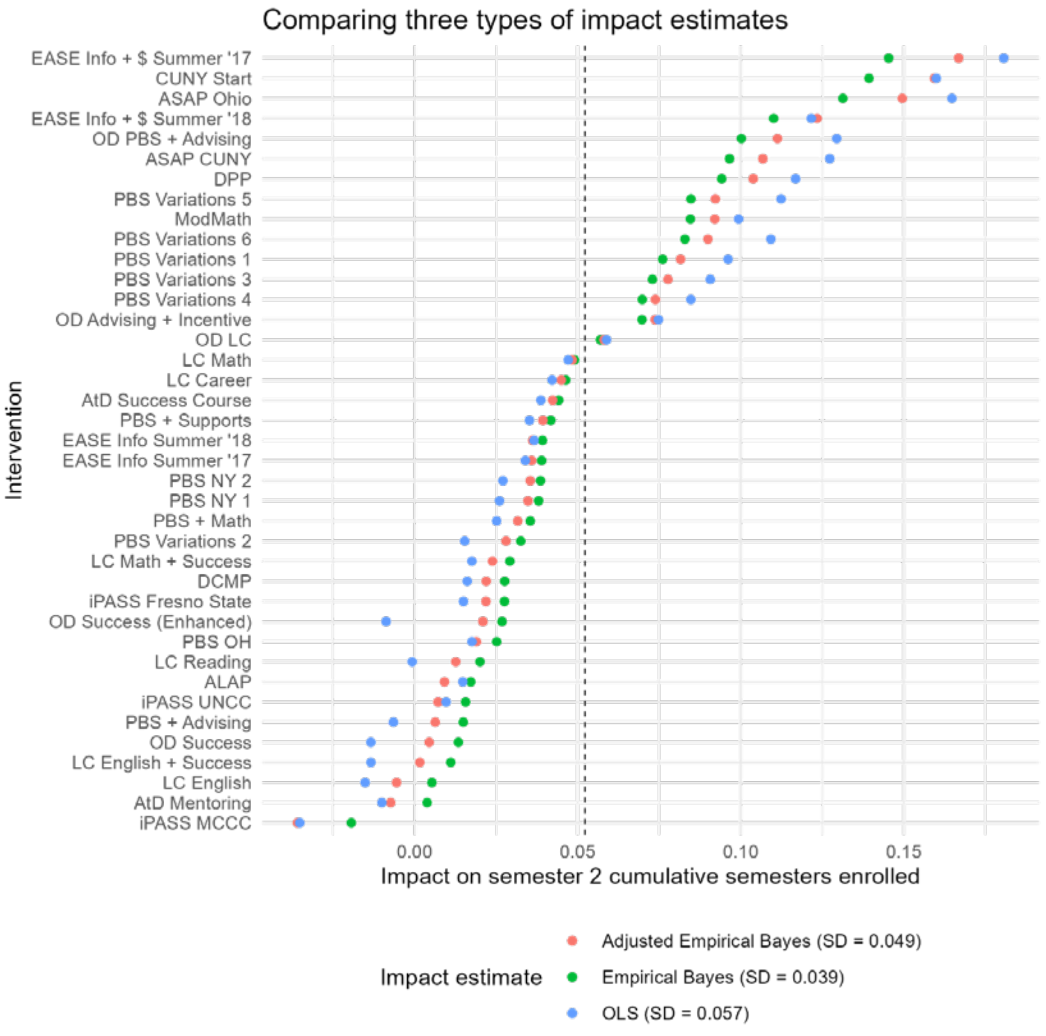
First, notice that the spread of the $\hat{B}_j^{Orig}$ is 16% larger than the spread of $\hat{B}_j^{AEB}$. Specifically, $SD\left(\hat{B}_j^{Orig}\right) = 0.057$, whereas the $SD\left(\hat{B}_j^{AEB}\right) = 0.049$. The latter, by construction, is equal to the estimate of τ, the standard deviation of the intervention-level distribution of average effects, from Equation 7. As shown in Equation 5, variation in $\hat{B}_j^{Orig}$ is expected to be larger than τ, owing to estimation error. The $\hat{B}_j^{AEB}$ aim to correct for this (and the $\hat{B}_j^{EB}$ slightly overcorrect for this in favor of other desirable statistical properties). Thus, the $\hat{B}_j^{AEB}$ may yield the most accurate representation of the spread of the distribution of true effects.

This is particularly relevant when planning a study and considering the MDTE. If a planned study has a MDTE of 0.125 semesters enrolled, examining the distribution of *estimated* effects ($\hat{B}_j^{Orig}$) would show that effects of at least 0.125 semesters enrolled have been observed in 5 out of 39 interventions. While a relatively large effect, researchers might argue that being in the top 12% of interventions seems feasible, depending on the intervention. Examining the distribution of adjusted empirical Bayes estimates ($\hat{B}_j^{AEB}$) might make telling that story more challenging. The adjusted empirical Bayes estimates suggests that an effect of 0.125 semesters enrolled has only occurred in 3 out of 39 interventions, implying that the intervention proposed for study would need to have a *true* effect in the top 6% of interventions in THE-RCT for the proposed study to be adequately powered. This could change researcher or funder decision-making.

Next, notice that the rank order of interventions' estimated effects occasionally changes, depending on the estimator. This is especially notable in evaluations where: (a) the original impact estimate is relatively more or less precise than the impact estimates for the other interventions (usually due to a relatively small/large sample size) and (b) the original impact estimate is far from the mean ($\hat{\beta}$) impact estimate. One notable example is the *EASE Info + $ Summer '18* intervention. While $\hat{B}_{EASE\ Info+\$Summer\ '18}^{Orig}$ is the sixth largest OLS impact estimate, $\hat{B}_{EASE\ Info+\$Summer\ '18}^{AEB}$ is the fourth largest adjusted

[17] To ensure estimator consistency across interventions and because many of the original studies did not estimate impacts on this outcome at this time point, the "original" impact estimates in Figure 1 were obtained from student-level data by estimating an ordinary least squares (OLS) regression with dependent variable cumulative semesters earned and independent variables:
(a) 0/1 indicators of students' random assignment block (typically defined by their cohort and/or college campus), and (b) interactions between a 0/1 indicator of students' treatment or control status and 0/1 identifiers of the intervention tested by the RCT that they were part of. The regression coefficients for these interaction terms are the "original" estimated effects of the intervention tested.

Figure 1. Intervention Effect Estimates Using Three Estimators



empirical Bayes impact estimate. Compared to the other evaluations at the top end of the distribution, *EASE Info + $ Summer '18* was a relatively large evaluation (total sample size around 3,500). Owing to the large sample size (and precise impact estimate), the difference between $\hat{B}^{Orig}_{EASE\ Info+\$Summer\ '18}$ and $\hat{B}^{AEB}_{EASE\ Info+\$Summer\ '18}$ is small and so *EASE Info + $ Summer '18*'s adjusted empirical Bayes impact estimate rank is better than its original impact estimate rank.

Such rank switching is probably a good thing—estimation error and related issues probably ought to result in shrunken expectations about the true effects of interventions with relatively impressive results coming from trials with imprecise impact estimates. Making such adjustments when interpreting findings from a new trial is prudent. The smallest trial in THE-RCT included 444 students, with all others including over 700 students. If a new trial with 150 students finds an estimated effect of 0.15 cumulative semesters enrolled through two semesters after random assignment, our best estimate probably should not be that this intervention is more effective at increasing cumulative semesters enrolled than nearly all other interventions in THE-RCT. By shrinking expectations, we can help combat "the winner's curse" and "promising trials bias" (Simpson, 2022; Sims et al., 2022). Appendix B provides a discussion of how to take the estimated effect of an intervention (and its associated standard error) and calculate an adjusted empirical Bayes impact estimate that can be located on the distribution of adjusted empirical Bayes impact estimates from this paper. An online tool associated with this paper allows users to make this adjustment easily, by plugging in an effect estimate and its standard error from a new RCT (go to https://www.mdrc.org/the-rct-empirical-benchmarks).

Figure 1 illustrated key points about our methodological approach using one outcome and time point as an example. We now turn to a broader discussion of the findings across major outcomes and time points.

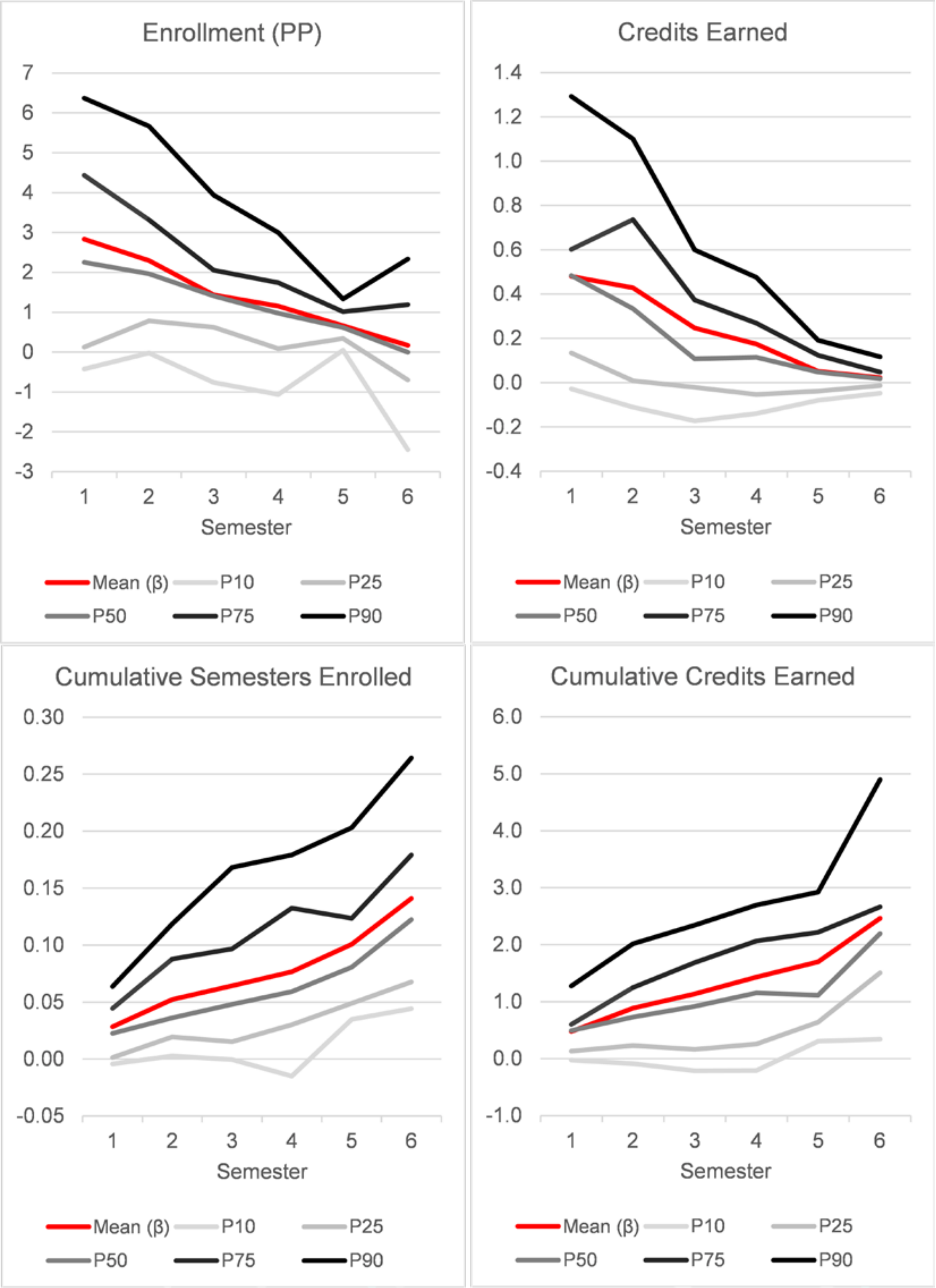
Estimated Distribution of True Impacts

Appendix Table A4 presents information on the estimated distribution of true effects (mean, standard deviation, and percentiles) across the 39 postsecondary interventions in THE-RCT, by outcome and by semester. The first two outcomes (enrollment and credits earned) are marginal, focused only on what happened in that semester. The other three outcomes (cumulative semesters enrolled, cumulative credits earned, and degree earned) are cumulative, such that earlier impacts carry forward. As discussed earlier, the distribution of effects beyond semester three or four should be interpreted with caution because these distributions are based on fewer studies and very likely biased upwards because studies with more promising short-term effects are more likely to have longer-term follow-up data. Figure 2 provides a visual summary of Appendix Table A4 for four outcomes and for six semesters after random assignment.

Consider first the effect distribution on credits earned through one semester. The mean of the distribution (β) is 0.48 credits. This implies that, on average, the interventions in THE-RCT had positive impacts on students' first semester credit accumulation. Next, is the estimate of true impact variation across interventions (τ). This cross-intervention standard deviation is estimated to be 0.55 credits. This estimate of τ is larger in magnitude than the mean impact (β), indicating substantial variation in the effectiveness of interventions under study on this outcome at this time point. Lastly, are estimates of the magnitude of effects at various points in the effect distribution—the 10th, 25th, 50th, 75th, and 90th percentiles are −0.03, 0.13, 0.48, 0.60, and 1.29, respectively.

Cutoffs to create rules of thumb for what is small, medium, or large are arbitrary, resulting in odd distinctions (e.g., an effect estimate of 0.59 credits might be considered medium and an effect estimate of 0.61 might be considered large, despite their being indistinguishable for practical or statistical reasons). Thus, it may be preferable to simply characterize the relative magnitude of an effect estimate in terms of its approximate

Figure 2. Points on the Estimated Distribution of Intervention Effects, By Outcome and Semester



Note. PP = percentage points. Percentiles are presented as PXX and are based on the adjusted empirical Bayes impact estimates. Sample sizes can be found in Table 4. See Appendix Table A4 for exact values.

percentile within the distribution of effects. Nevertheless, some people find rules of thumb helpful, so one might consider effects on credits earned in semester one below 0.13 (25th percentile) to be relatively small, between 0.13 and 0.60 credits (75th percentile) to be medium-sized, and above 0.60 credits to be relatively large.

When looking across outcomes and semesters, a few patterns emerge:

- **For marginal outcomes, downward trends for means, and decreasing variability over time.**

The mean (across interventions) effect on marginal enrollment and credits earned decreases over time. For example, the mean effect on credits earned starts at 0.48 credits in semester 1, and decreases to 0.43, 0.25, 0.17, 0.05, and 0.02 credits in semesters 2–6, respectively. This downward trend holds for enrollment. Similarly, the spread of the intervention-level distribution of effects is widest in semester 1 and decreases over time. These trends suggest that a lot of the action, with respect to intervention effects, occurs early on. Consequently, in the short-term, an intervention's effects on enrollment and credits earned must be larger in absolute magnitude to be considered large relative to the effects of other interventions.

- **For cumulative outcomes, upward trends, and increased variability over time.**

The mean (across interventions) effect on cumulative number of semesters enrolled and cumulative credits earned increases over time. For example, the mean effect on cumulative credits earned starts at 0.47 credits in semester 1, and increases to 0.89, 1.14, 1.43, 1.70, and 2.46 credits, in semesters 2–6, respectively. This pattern holds for cumulative enrollment. Similarly, the spread of the intervention-level distribution of effects is smallest in semester 1 and increases over time. These trends show that the effects of some of the more effective interventions continue to grow throughout the first three years after random assignment. Consequently, over time, an intervention's effects on cumulative enrollment and cumulative credits earned must be larger in absolute magnitude to be considered large relative to the effects of other interventions.

Because degree completion is a longer-term outcome, the number of studies with follow-up data is smaller than for other outcomes. Based on the subset of studies with available data (20 to 23 interventions, depending on the semester), the mean effect size is 0.9 percentage points by semester 4, 1.6 percentage points by semester 5, and 1.7 percentage points by semester 6. As noted earlier, these average effects are likely biased upwards due to follow-up selection bias. One notable aspect of the degree findings is that two outlier interventions drive a lot of the action. The original effect estimates from these studies are approximately 16 and 18 percentage points. No other study's original effect estimate is above 4 percentage points. Thus, benchmarking degree impacts based on the evaluations in THE-RCT is limited. It seems reasonable to consider any positive effect on degree completion to be an impressive feat, with effects larger than 5 percentage points being notable.

Discussion

We break our discussion into a few parts: planning, interpretation, and other remarks.

Planning

The findings in this paper can be used to plan the sample size for future evaluations of CC interventions. For example, when planning a study of an intervention that is lighter-touch (perhaps a single-component intervention), researchers may want to choose a sample size that will make it possible to detect effects at the lower end of the distribution of effects presented in this paper (for example, 0.23 cumulative credits earned through one year). Whereas for more comprehensive interventions, which tend to have larger effects (Weiss & Bloom, 2022; Weiss et al., 2022), a larger minimum detectable true effect may be sufficient, depending on the goal of the study. Importantly, other design parameters are necessary to calculate the minimum detectable true effect of an intervention—for guidance specific to community college evaluations, see Somers et al. (2023).

Notably, researchers should be mindful of the outcome(s) of interest and the timing of measuring those outcomes—this has implications for what sized effects might realistically be achieved. For example, the size of the effect an intervention might reasonably achieve on marginal credits earned in semester 2 is quite different than the size of the effect an intervention might achieve on cumulative credits earned through semester 4. Thus, MDTE calculations need to be specific with respect to outcome and timing.

Interpretation

The findings presented in this paper begin to provide empirical benchmarks to help researchers and policymakers interpret effect estimates from evaluations of CC interventions. Returning to the example of the DPP intervention highlighted in the introduction, recall that the estimated effect of DPP was 1.8 credits earned after two semesters (Ratledge et al., 2021, Appendix Table B3). This results in an adjusted empirical Bayes impact estimate of 1.6 credits earned, which is about 0.70 standard deviations above the mean effect, and at the 83rd percentile in the distribution of effects after two semesters.[18] Thus, the effect of the DPP intervention is quite large compared to effects of other evaluated interventions.

[18] The Detroit Promise Path evaluation is part of THE-RCT. This result is thus based on the adjusted empirical Bayes estimate from FIRC. Appendix B describes how this would be calculated for a new study.

As noted earlier, the interventions included in our analysis vary in terms of their components and duration (see Table 1). For this reason, in theory, it may be appropriate for researchers to compare the effect of their study to the findings from prior studies of similar interventions. To this end, study-level estimates of intervention effects (OLS, empirical Bayes, and adjusted empirical Bayes) for the 39 interventions are available in a public-use dataset created for this paper (https://www.mdrc.org/the-rct-empirical-benchmarks). The dataset includes estimated effects for each of the 39 studies in the analysis, by semester, for all outcomes (including additional outcomes like credits attempted and credits earned, for total, college-level, and developmental credits), allowing researchers to look at estimated effects for narrower categories of interventions as well as additional outcomes. As more CC studies are conducted, it may be possible to look at effect sizes for different populations, and a broader array of interventions.

Researchers of CC studies can also consider using other types of benchmarks to interpret their findings. One of the limitations of using effects from prior studies as benchmarks is that while they inform what is realistically attainable, they do not necessarily inform what effects are practically meaningful to decision-makers. In K-12 research, several other approaches have been offered for interpreting the practical meaningfulness of effect sizes. This includes comparing a study's effects to "normative expectations for change or growth" (e.g., comparing a study's effect to the typical growth made by a student during the year) and policy-relevant performance gaps (e.g., comparing a study's effect to the outcomes gap between students from families with low versus higher income; Baird & Pane, 2019; Bloom et al., 2008; Hill et al., 2008; Konstantopoulos & Hedges, 2008; Kraft, 2020; Lipsey et al., 2012; Wolf & Harbatkin, 2022).

These approaches could also be used to interpret the practical meaningfulness of findings from CC studies. With respect to normative expectations for growth, the magnitude of a program's effect on college-level credit accumulation could be characterized relative to normal academic progress towards a degree over various time periods. For example, for students who first enrolled in a public CC in 2012, average credit accumulation over one year nationally was 13.2 credits.[19] Therefore, if an intervention's estimated effect on credit accumulation through two semesters is 3.0 credits, then the intervention could be said to increase credit accumulation by 25% (3.0/12.0) of the national average credit accumulation.

Similarly, comparing estimated effects to inequality in academic outcomes across relevant populations could also be used to characterize impacts. For example, nationally, among students who first enrolled in a public CC in 2017, there is racial inequality in three-year graduation rates of 10.5 percentage points between Hispanic men and White men (U.S. Department of Education, 2021). Thus, an intervention targeting Latino males with a 1.4 percentage point effect on three-year graduation rates could

[19] Based on tables from the 2012 cohort of the Beginning Postsecondary Students (BPS) study, extracted using the NCES DataLab tool (https://nces.ed.gov/datalab).

be characterized as reducing racial inequality in graduation rates among Hispanic men and White men by about 13%.

When using normative benchmarks (whether progress towards a degree or racial inequality in academic outcomes), an important consideration is what the reference population should be, which in turn depends on how the findings will be used. The previous examples were based on a national reference population, which may be relevant if the goal is to understand the potential of an intervention to address racial inequality if it were scaled to additional CCs across the country. Information about postsecondary outcomes and achievement gaps for nationally representative samples are readily available from public data sources like the Integrated Postsecondary Education Data System (IPEDS) and the Beginning Postsecondary Students (BPS) study conducted by the National Center for Education Statistics (NCES).[20] However, for policymakers or practitioners working at the state or institutional level, benchmarks based on a local normative population could be more appropriate. For example, if an intervention is "home grown" by a state, then state policymakers may find it most useful to interpret the effect sizes from a study relative to the racial inequality in academic outcomes in their state. Similarly, at the colleges that are implementing the intervention being evaluated, administrators may want to know by how much the intervention reduces racial inequality in student outcomes for students at their institution, or even more specifically, for the subset of students who were eligible to receive the intervention. Information on a college's racial inequality in academic outcomes can often be obtained directly from the institution, and gaps for participating students can be estimated using the study's control group.

For policymakers and practitioners, practical considerations related to an intervention's implementation can be as important as its effectiveness. Hence, other powerful approaches for contextualizing a study's effects are to discuss the intervention's cost-effectiveness (impact per dollar spent) and its scalability to different contexts (Kraft, 2020). Notably, 19 of the interventions in THE-RCT are part of MDRC's *Intervention Return on Investment (ROI) Tool for Community Colleges* (see https://www.mdrc.org/intervention-roi-tool), and thus estimates of the direct costs of these interventions are publicly available. Careful thought would be required to pull this information together to create cost-effectiveness benchmarks—an important area for future investigation that we hope to pursue.

[20] Descriptive statistics on postsecondary outcomes (e.g., enrollment, credit accumulation, credential attainment) based on the BPS and IPEDS are provided in NCES reports and the Digest of Education Statistics. Researchers can also create their own descriptive tables for specific subgroups using the NCES DataLab tool (https://nces.ed.gov/datalab).

Other Remarks

Intent-to-Treat and Treatment-on-the-Treated

In most of the studies in THE-RCT eligible students who were interested in participating in the intervention were recruited and consented to participate in the evaluation. Consequently, intervention take-up rates were high (typically over 70%) and there is limited difference between the magnitude of effect of the intent-to-treat (ITT) compared with the effect of the treatment-on-the-treated (TOT), or the local average treatment effect (LATE; Angrist et al., 1996; Bloom, 1984). Thus, it is probably inappropriate to compare the ITT effects from an evaluation that randomizes all eligible students "behind-the-scenes" and yields a low take-up rate to the distribution of ITT effects presented in the present paper. Comparing TOT or LATE estimates from a study with low take-up rates may be more appropriate; however, even this should be done with caution owing to the often-inflated standard errors (when take-up is low) and the fact that, despite high take-up rates, the effect estimates in the present paper are indeed ITT and would need to be inflated to make them more comparable to TOT estimates. Perhaps this is an area of future work.

A Warning About Standardized Effect Sizes

The standardized mean effect size is the program-control group difference in mean outcomes divided by the standard deviation of the outcome (typically for the control group or pooled within research groups). When planning an evaluation, it is common for researchers to calculate the minimum detectable true effect size (MDTES) in standardized units, rather than the minimum detectable true effect (MDTE) in natural units (e.g., credits earned, percentage points for enrollment or degrees earned). The same is sometimes done when describing findings—intervention effect estimates may be presented in "standardized" effect size units. This is especially common when combining effect estimates for the purpose of meta-analysis. We offer caution for those tempted to do so.

First, standardized effect sizes may result in an unnecessary lack of transparency. They are common in K-12 research because test scores are often the outcome measure of interest and test scores are on arbitrary scales, thus some form of standardization is necessary for interpretation. In contrast, in CC research most outcomes have meaning in their natural units—graduation rates, enrollment rates, credits earned. Thus, a very strong reason is needed to justify converting effect estimates from a unit with natural meaning to one that is difficult to interpret and tethered to the amount of variation in the outcome among a specific sample of individuals.

Second, as we show in this paper, the distribution of effects across interventions can vary over time, with changing distributional means and variances. Moreover, as shown in Somers et al. (2023), the standard deviation of the outcome also changes over time, among outcomes, and across interventions. Combining these facts, it may be quite difficult to make sense of comparisons of standardized effect sizes across time, outcomes, or evaluations. The cleanest comparisons or pooling may involve the same (or very similar) outcomes measured at the same time point.

Methodological Next Steps for Empirical Benchmarks Researchers

In addition to providing empirical benchmarks for planning evaluations and interpreting effect estimates from evaluations in a new context (CCs), this paper also offers some methodological advances. By using random effects models and adjusted empirical Bayes impact estimates, we move closer to providing empirical benchmarks that represent the intervention-level distribution of *true* effects from past evaluations, rather than the intervention-level distribution of *estimated* effects from past evaluations.

This advance can be further improved upon. Analyses for each outcome at each time point were conducted independently, despite known correlations among impacts over time and across outcomes. Pooling data could be beneficial, especially given the lack of precision when estimating key parameters of interest. Specifically, the kinks in Figure 2 and some oddities in Appendix Table A4 likely illustrate the challenge of separating signal from noise when estimating τ with a limited number of studies. This challenge carries through to the adjusted empirical Bayes impact estimates and thus the percentiles of the intervention-level distribution of effects. This noise could be smoothed by pooling the data over time within outcome (at a minimum), and by forcing structure on the estimates of the τ's, such as assuming a plausible functional form over time (linearity or curvilinearity). Such pooling might also mitigate some of the concern around follow-up selection bias, since estimates of β and τ at later time points would be estimated, in part, based on data from the earlier time points with more complete data.

A Plea to Postsecondary Researchers and Funders

This article is one in a series of papers that capitalize on the unique dataset known as THE-RCT (for example, see Bailey & Weiss, 2022; Somers et al., 2023; Weiss & Bloom, 2022; Weiss et al., 2021, 2022). In addition to promoting increased learning through open and transparent data sharing, THE-RCT facilitates cross-study knowledge building. This was supported by the creation of core outcome measures, available semesterly, across all studies in THE-RCT.

To the extent that postsecondary researchers and funders value this type of cross-study learning, it is imperative to the field that we: (a) agree upon core outcome measures that are examined across studies (even if the outcome is not the primary outcome of the study) and (b) present impact estimates and associated standard errors (even if only in appendices), by semester for these core outcomes.

Conclusion

This paper provides an important first step in helping planners and potential funders of evaluations of CC interventions consider whether a proposed study is adequately powered to detect realistically achievable effects and supporting consumers of CC research interpret the magnitude of effects from rigorous evaluations. We hope that others will expand upon this work to further the field.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D190025 to MDRC. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors thank Howard Bloom for his careful review and drafting of Appendix B.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455. https://doi.org/10.2307/2291629

Bailey, D., & Weiss, M. J. (2022, August 18). Do meta-analyses oversell the longer-term effects of programs? (Part 1). *Detecting follow-up selection bias in studies of postsecondary education programs*. MDRC. https://www.mdrc.org/publication/do-meta-analyses-oversell-longer-term-effects-programs-part-1

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. *Educational Researcher*, *48*(4), 217–228. https://doi.org/10.3102/0013189x19848729

Bloom, H. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review*, *19*(5), 547–556. https://doi.org/10.1177/0193841X9501900504

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*(2), 225–246. https://doi.org/10.1177/0193841X8400800205

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*(4), 289–328. https://doi.org/10.1080/19345740802400072

Bloom, H. S., Raudenbush, S., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842. https://doi.org/10.1080/19345747.2016.1264518

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292. https://doi.org/10.3102/0013189X16656615

Diamond, J., Weiss, M. J., Hill, C., Slaughter, A., & Dai, S. (2021). *MDRC's The Higher Education Randomized Controlled Trials Restricted Access File (THE-RCT RAF), United States, 2003–2019* Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR37932.v1

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217. https://doi.org/10.1037//1082-989X.6.3.203

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Integrated Postsecondary Education Data System (IPEDS) Trend Generator. (2020). *Number of students enrolled in postsecondary institutions in the fall, by sector of institution and level of student: 2020*. Retrieved from https://nces.ed.gov/ipeds/TrendGenerator/app/build-table/2/3?rid=1&cid=14

Integrated Postsecondary Education Data System (IPEDS) Trend Generator. (2021). *Graduation rate within 150% of normal time at 2-year postsecondary institutions*. Retrieved from https://nces.ed.gov/ipeds/TrendGenerator/app/answer/7/21

Konstantopoulos, S., & Hedges, L. V. (2008). How large an effect can we expect from school reforms? *Teachers College Record*, *110*(8), 1611–1638. https://doi.org/10.1177/016146810811000803

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kraft, M. A. (2023). The effect-size benchmark that matters most: Education interventions often fail. *Educational Researcher*, *52*(3), 183–187. https://doi.org/10.3102/0013189x231155154

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013–3000). National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Ratledge, A., Sommo, C., Cullinan, D., O'Donoghue, R., Lepe, M., & Camo-Biogradlija, J. (2021). *Motor City Momentum—Three years of the Detroit Promise Path Program for community college students*. MDRC. https://www.mdrc.org/sites/default/files/DetroitPromis_Path-Final.pdf

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475–499. https://doi.org/10.1177/1098214015600515

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2^{nd} ed.). Sage Publications.

Simpson, A. (2022). A recipe for disappointment: Policy, effect size, and the winner's curse. *Journal of Research on Educational Effectiveness*. Advance online publication. https://doi.org/10.1080/19345747.2022.2066588

Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2022). Quantifying "promising trials bias" in randomized controlled trials in education. *Journal of Research on Educational Effectiveness*. Advance online publication. https://doi.org/10.1080/19345747.2022.2090470

Somers, M.-A., Weiss, M. J., & Hill, C. (2023). Design parameters for planning the sample size of individual-level RCTs in community colleges. *Evaluation Review*, *47*(4), 599–629. https://doi.org/10.1177/0193841X221121236

Spybrook, J., Zhang, Q., Kelcey, B., & Dong, N. (2020). Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine what works, for whom, and under what conditions. *Educational Evaluation and Policy Analysis*, *42*(3), 354–374. https://doi.org/10.3102/0162373720929018

Torgerson, C. J., Torgerson, D. J., Birks, Y. F., & Porthouse, J. (2005). A comparison of randomised controlled trials in health and education. *British Educational Research Journal*, *31*(6), 761–785. https://doi.org/10.1080/01411920500314919

U.S. Department of Education. (2021). *Digest of education statistics: 2021* (Table 326.20). National Center for Education Statistics. https://nces.ed.gov/programs/digest/d21/tables/dt21_326.20.asp

Weiss, M. J., & Bloom, H. (2022). *What works for community college students? A brief synthesis of 20 years of MDRC's randomized controlled trials*. MDRC. https://www.mdrc.org/sites/default/files/THE-RCT_Synthesis_Brief.pdf

Weiss, M. J., Bloom, H. S., & Singh, K. (2022). What 20 years of MDRC RCTs suggest about predictive relationships between intervention features and intervention impacts for community college students. *Educational Evaluation and Policy Analysis*. Advance online publication. https://doi.org/10.3102/01623737221139493

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876. https://doi.org/10.1080/19345747.2017.1300719

Weiss, M. J., Unterman, R., & Biedzio, D. (2021). *What happens after the program ends? A synthesis of post-program effects in higher education*. MDRC. https://www.mdrc.org/sites/default/files/THE-RCT_Fade-Out_IF_Final.pdf

Wolf, B., & Harbatkin, E. (2023). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, *16*(1), 134–161. https://doi.org/10.1080/19345747.2022.2071364

Appendix Table A1

THE-RCT Study Abbreviations, Study Names, and References

Study Abbreviation	Study Name	References
ALAP	Aid Like a Paycheck	Weissman, E., Cerna, O., Cullinan, D., & Baldiga, A. (2017). *Aligning aid with enrollment: Interim findings on Aid Like a Paycheck*. MDRC. https://www.mdrc.org/sites/default/files/ALAP_Interim_Report_2017.pdf
		Weissman, E., Cerna, O., & Cullinan, D. (2019). *Incremental disbursements of student financial aid: Final report on Aid Like a Paycheck*. MDRC. https://www.mdrc.org/sites/default/files/ALAP_2019_FINAL_rev.pdf
ASAP CUNY	Accelerated Study in Associate Programs—City University of New York	Scrivener, S., Weiss, M. J., Ratledge, A., Rudd, T., Sommo, C., & Fresques, H. (2015). *Doubling graduation rates: Three-year effects of CUNY's Accelerated Study in Associate Programs (ASAP) for developmental education students*. MDRC. https://www.mdrc.org/sites/default/files/doubling_graduation_rates_fr.pdf
		Weiss, M., Ratledge, A., Sommo, C., & Gupta, H. (2019). Supporting community college students from start to degree completion: Long-term evidence from a randomized trial of CUNY'S ASAP. *American Economic Journal: Applied Economics*, *11*(3), 253–297. https://doi.org/10.1257/app.20170430
		Azurdia, G., & Galkin, K. (2020). *An eight-year cost analysis from a randomized controlled trial of CUNY's Accelerated Study in Associate Programs*. MDRC. https://www.mdrc.org/sites/default/files/ASAP_Cost_Working_Paper_final.pdf
ASAP Ohio	Ohio Accelerated Study in Associate's Programs—Ohio Replication	Sommo, C., Cullinan, D., & Manno, M. (2018, December). *Doubling graduation rates in a new state: Two-year findings from the ASAP Ohio Demonstration*. MDRC. https://www.mdrc.org/sites/default/files/ASAP_brief_2018_Final.pdf
		Miller, C., & Weiss, M. J. (2021). *Increasing community college graduation rates: A synthesis of findings on the ASAP model from six colleges across two states*. MDRC. https://www.mdrc.org/sites/default/files/ASAP_OH_3yr_Impact_Report_1.pdf

THE-RCT Study Abbreviations, Study Names, and References (*continued*)

Study Abbreviation	Study Name	References
AtD Mentoring	Achieving the Dream—Beacon Mentoring Program	Visher, M., Butcher, K. F., & Cerna, O. S. (2010). *Guiding developmental math students to campus services: An impact evaluation of the Beacon Program at South Texas College*. MDRC & Achieving the Dream. https://www.mdrc.org/sites/default/files/full_382.pdf
AtD Success Course	Achieving the Dream—Student Success Course	Rutschow, E. Z., Cullinan, D., & Welbeck, R. (2012). *Keeping students on course: An impact study of a student success course at Guilford Technical Community College*. MDRC & Achieving the Dream. https://www.mdrc.org/sites/default/files/Keeping%20Students%20on%20Course%20Full%20Report.pdf
CUNY Start	CUNY Start	Scrivener, S., Gupta, H., Weiss, M. J., Cohen, B., Scott Cormier, M., & Brathwaite, J. (2018). *Becoming college-ready: Early findings from a CUNY Start evaluation*. MDRC & Community College Research Center. https://www.mdrc.org/sites/default/files/CUNY_START_Interim_Report_FINAL_0.pdf Weiss, M. J., Scrivener, S., Slaughter, A., & Cohen, B. (2021). An on-ramp to student success: A randomized controlled trial evaluation of a developmental education reform at the City University of New York. *Educational Evaluation and Policy Analysis*, *43*(4), 555–586. https://doi.org/10.3102/01623737211008901
DCMP	Dana Center Math Pathways	Rutschow, E. Z. (2018). *Making it through: Interim findings on developmental students' progress to college math with the Dana Center Mathematics Pathways*. Center for the Analysis of Postsecondary Readiness. https://www.mdrc.org/sites/default/files/DCMP-InterimFindings.pdf Rutschow, E. Z., Sepanik, S., Deitch, V., Raufman, J., Dukes, D., & Moussa, A. (2019). *Gaining ground: Findings from the Dana Center Mathematics Pathways impact study*. Center for the Analysis of Postsecondary Readiness. https://www.mdrc.org/sites/default/files/DCMP_Final_Report_2019.pdf

THE-RCT Study Abbreviations, Study Names, and References (*continued*)

Study Abbreviation	Study Name	References
DPP	Detroit Promise Path	Ratledge, A., & Vasquez, A. (2018, May). *Learning from success: The Detroit Promise Path*. MDRC. https://www.mdrc.org/sites/default/files/Detroit_Promise_Path_Issue_Focus.pdf Ratledge, A., O'Donoghue, R., Cullinan, D., & Camo-Biogradlija, J. (2019). *A path from access to success: Interim findings from the Detroit Promise Path evaluation*. MDRC. https://www.mdrc.org/sites/default/files/Detroit_Promise_Path_Report-Final_0.pdf
EASE*	Encouraging Additional Summer Enrollment	Headlam, C., Anzelone, C., & Weiss, M. J. (2018, July). *Making summer pay off: Using behavioral science to encourage postsecondary summer enrollment*. Center for Applied Behavioral Sciences at MDRC. https://www.mdrc.org/sites/default/files/EASE_Phase_1_Brief_Final_Web.pdf Weiss, M. (2019, February). *How can community colleges increase student use of year-round Pell Grants? Two proven strategies to boost summer enrollment*. Center for Applied Behavioral Sciences at MDRC. https://www.mdrc.org/sites/default/files/EASE_Brief_Phase%202_Final2.pdf Headlam, C., Cohen, B., & Reiman, K. (2020, April). *EASE handbook for community colleges: Encouraging summer enrollment*. Center for Applied Behavioral Sciences at MDRC. https://www.mdrc.org/sites/default/files/EASE_Practitioner_Guide_2020_0.pdf
iPASS Fresno State	Integrated Planning and Advising for Student Success—California State University Fresno State	Mayer, A., Kalamkarian, H. S., Cohen, B., Pellegrino, L., Boynton, M., & Yang, E. (2019). *Integrating technology and advising: Studying enhancements to colleges' iPASS practices*. MDRC. https://www.mdrc.org/sites/default/files/iPASS_Interim_Report.pdf
iPASS MCCC	Integrated Planning and Advising for Student Success—Montgomery County Community College	See Mayer et al. (2019).

THE-RCT Study Abbreviations, Study Names, and References (*continued*)

Study Abbreviation	Study Name	References
iPASS UNCC	Integrated Planning and Advising for Student Success—University of North Carolina at Charlotte	See Mayer et al. (2019).
LC Career	Learning Communities—Career Focused at Kingsborough Community College	Visher, M., & Teres, J. (2011). *Breaking new ground: An impact study of career-focused learning communities at Kingsborough Community College*. National Center for Postsecondary Research. https://www.mdrc.org/sites/default/files/full_382.pdf Visher, M. G., Weiss, M. J., Weissman, E., Rudd, T., & Wathington, H. D. (2012). *The effects of learning communities for students in developmental education*. National Center for Postsecondary Research. https://www.mdrc.org/sites/default/files/LC%20A%20Synthesis%20of%20Findings%20FR.pdf
LC English	Learning Communities—Developmental English	Weissman, E., Cullinan, D., Cerna, O., Safran, S., & Richman, P. (2012). *Learning communities for students in developmental English: Impact studies at Merced College and the Community College of Baltimore County*. National Center for Postsecondary Research. https://www.mdrc.org/sites/default/files/full_422.pdf Weiss, M. J., Visher, M. G., Weissman, E., & Wathington, H. (2015). The impact of learning communities for students in developmental education: A synthesis of findings from randomized trials at six community colleges. *Educational Evaluation and Policy Analysis*, *37*(4), 520–541. https://doi.org/10.3102/0162373714563307
LC English + Success	Learning Communities—Developmental English + Success Course	See Weissman, E., Cullinan, D., Cerna, O., Safran, S., & Richman, P. (2012). See Weiss, M. J., Visher, M. G., Weissman, E., & Wathington, H. (2015).
LC Math	Learning Communities—Developmental Math at Queensborough and Houston Community Colleges	Weissman, E., Butcher, K. F., Schneider, E., Teres, J., Collado, H., & Greenberg, D. (2011). *Learning communities for students in developmental math: Impact studies at Queensborough and Houston Community Colleges*. National Center for Postsecondary Research. https://www.mdrc.org/sites/default/files/full_423.pdf See Weiss, M. J., Visher, M. G., Weissman, E., & Wathington, H. (2015).

THE-RCT Study Abbreviations, Study Names, and References (*continued*)

Study Abbreviation	Study Name	References
LC Math + Success	Learning Communities—Developmental Math + Success Course	See Weissman, E., Butcher, K. F., Schneider, E., Teres, J., Collado, H., & Greenberg, D. (2011). See Weiss, M. J., Visher, M. G., Weissman, E., & Wathington, H. (2015).
LC Reading	Learning Communities—Developmental Reading at Hillsborough Community College	Weiss, M., Visher, M., & Wathington, H. (2010). *Learning communities for students in developmental reading: An impact study at Hillsborough Community College*. National Center for Postsecondary Research. https://www.mdrc.org/sites/default/files/full_424.pdf See Weiss, M. J., Visher, M. G., Weissman, E., & Wathington, H. (2015).
ModMath	Modularized, Computer-Assisted Developmental Math	Gardenhire, A., Diamond, J., Headlam, C., & Weiss, M. J. (2016). *At their own pace: Interim findings from an evaluation of a computer-assisted, modular approach to developmental math*. MDRC. https://www.mdrc.org/sites/default/files/ModMath%20Report%202016.pdf Weiss, M. J., & Headlam, C. (2019). A randomized controlled trial of a modularized, computer-assisted, self-paced approach to developmental math. *Journal of Research on Educational Effectiveness*, *12(*3), 484–513. https://doi.org/10.1080/19345747.2019.1631419
OD Advising + Incentive	Opening Doors—Advising + Financial Incentive	Scrivener, S., & Weiss, M. J. (2009). *More guidance, better results? Three-year effects of an enhanced student services program at two community colleges*. MDRC. https://www.mdrc.org/sites/default/files/full_450.pdf Scrivener, S., & Coghlan, E. (2011, March). *Opening doors to student success: A synthesis of findings from an evaluation at six community colleges*. MDRC. https://www.mdrc.org/sites/default/files/policybrief_27.pdf

THE-RCT Study Abbreviations, Study Names, and References (*continued*)

Study Abbreviation	Study Name	References
OD LC	Opening Doors—Comprehensive Learning Community	Scrivener, S., Bloom, D., LeBlanc, A., Paxson, C., Rouse, C. E., & Sommo, C. (2008). *A good start: Two-year effects of a freshmen learning community program at Kingsborough Community College*. MDRC. https://www.mdrc.org/sites/default/files/A%20Good%20Start.pdf Weiss, M., Mayer, A., Cullinan, D., Ratledge, A., Sommo, C., & Diamond, J. (2015). A random assignment evaluation of learning communities at Kingsborough Community College: Seven years later. *Journal of Research on Educational Effectiveness*, *8*(2), 189–217. https://doi.org/10.1080/19345747.2014.946634 See Scrivener, S., & Coghlan, E. (2011).
OD PBS + Advising	Opening Doors—Performance Based Scholarship + Advising	Richburg-Hayes, L., Brock, T., LeBlanc, A., Paxson, C., Rouse, C. E., & Barrow, L. (2009). *Rewarding persistence: Effects of a performance-based scholarship program for low-income parents*. MDRC. https://www.mdrc.org/sites/default/files/rewarding_persistence_fr.pdf Patel, R., Richburg-Hayes, L., de la Campa, E., & Rudd, T. (2013, August). *Performance-based scholarships: What have we learned? Interim findings from the PBS Demonstration*. MDRC. https://www.mdrc.org/sites/default/files/pbs_what_have_we_learned.pdf See Scrivener, S., & Coghlan, E. (2011).
OD Success	Opening Doors—College Success Course + Centers	Scrivener, S., Sommo, C., & Collado, H. (2009). *Getting back on track: Effects of a community college program for probationary students*. MDRC. https://www.mdrc.org/sites/default/files/full_379.pdf See Scrivener, S., & Coghlan, E. (2011).
OD Success (Enhanced)	Opening Doors—College Success Course + Centers (Enhanced)	See Scrivener, S., Sommo, C., & Collado, H. (2009). Weiss, M., Brock, T., Sommo, C., Rudd, T., & Turner, M. C. (2011). *Serving community college students on probation: Four-year findings from Chaffey College's Opening Doors Program*. MDRC. https://www.mdrc.org/sites/default/files/full_506.pdf See Scrivener, S., & Coghlan, E. (2011).

THE-RCT Study Abbreviations, Study Names, and References (*continued*)

Study Abbreviation	Study Name	References
PBS + Advising	Performance Based Scholarships + Advising—New Mexico	Binder, M., Krause, K., Miller, C., & Cerna, O. (2015). *Providing incentives for timely progress toward earning a college degree*. MDRC Working Paper. https://www.mdrc.org/sites/default/files/PBS_New-Mexico.pdf See Patel, R., Richburg-Hayes, L., de la Campa, E., & Rudd, T. (2013). Mayer, A. K., Patel, R., Rudd, T., & Ratledge, A. (2015). *Designing scholarships to improve college success*. MDRC. https://www.mdrc.org/sites/default/files/designing_scholarships_FR.pdf
PBS + Math	Performance Based Scholarships + Math Lab—Florida	Sommo, C., Boynton, M., Collado, H., Diamond, J., Gardenhire, A., Ratledge, A., Rudd, T., & Weiss, M. J. (2014). *Mapping success: Performance-based scholarships, student services, and developmental math at Hillsborough Community College*. MDRC. https://www.mdrc.org/sites/default/files/PBS-HCC%20 2014%20Full%20Report.pdf See Patel, R., Richburg-Hayes, L., de la Campa, E., & Rudd, T. (2013). See Mayer, A. K., Patel, R., Rudd, T., & Ratledge, A. (2015).
PBS + Supports	Performance Based Scholarships + Supports—Arizona	Patel, R., & Valenzuela, I. (2013). *Moving forward: Early findings from the performance-based scholarship demonstration in Arizona*. MDRC. https://www.mdrc.org/sites/default/files/Moving_Forward_FR_0.pdf See Patel, R., Richburg-Hayes, L., de la Campa, E., & Rudd, T. (2013). See Mayer, A. K., Patel, R., Rudd, T., & Ratledge, A. (2015).
PBS NY*	Performance Based Scholarships—New York	Richburg-Hayes, L., Sommo, C., & Welbeck, R. (2011). *Promoting full-time attendance among adults in community college: Early impacts from the performance-based scholarship demonstration in New York*. MDRC. https://www.mdrc.org/sites/default/files/full_480.pdf See Patel, R., Richburg-Hayes, L., de la Campa, E., & Rudd, T. (2013). See Mayer, A. K., Patel, R., Rudd, T., & Ratledge, A. (2015).

THE-RCT Study Abbreviations, Study Names, and References (*continued*)

Study Abbreviation	Study Name	References
PBS OH	Performance Based Scholarships—Ohio	Mayer, A., Patel, R., & Gutierrez, M. (2015). *Four-year effects on degree receipt and employment outcomes from a performance-based scholarship program in Ohio*. MDRC Working Paper. http://www.mdrc.org/sites/default/files/Four-Year_Effects_on_Degree_Receipt_0.pdf
		See Patel, R., Richburg-Hayes, L., de la Campa, E., & Rudd, T. (2013).
		See Mayer, A. K., Patel, R., Rudd, T., & Ratledge, A. (2015).
PBS Variations*	Performance Based Scholarships, Varying Amounts—California	Richburg-Hayes, L., Patel, R., Brock, T., de la Campa, E., Rudd, T., & Valenzuela, I. (2015). *Providing more cash for college: Interim findings from the performance-based scholarship demonstration in California*. MDRC. https://www.mdrc.org/sites/default/files/Providing_More_Cash_FR.pdf
		See Patel, R., Richburg-Hayes, L., de la Campa, E., & Rudd, T. (2013).
		See Mayer, A. K., Patel, R., Rudd, T., & Ratledge, A. (2015).

Note. Studies examining the effects of more than one intervention are indicated with an asterisk. They were collapsed into a single row because all research groups are described within the same references.

Appendix Table A2

Description of Codes for Intervention Features (Cumulative Through Year 1)

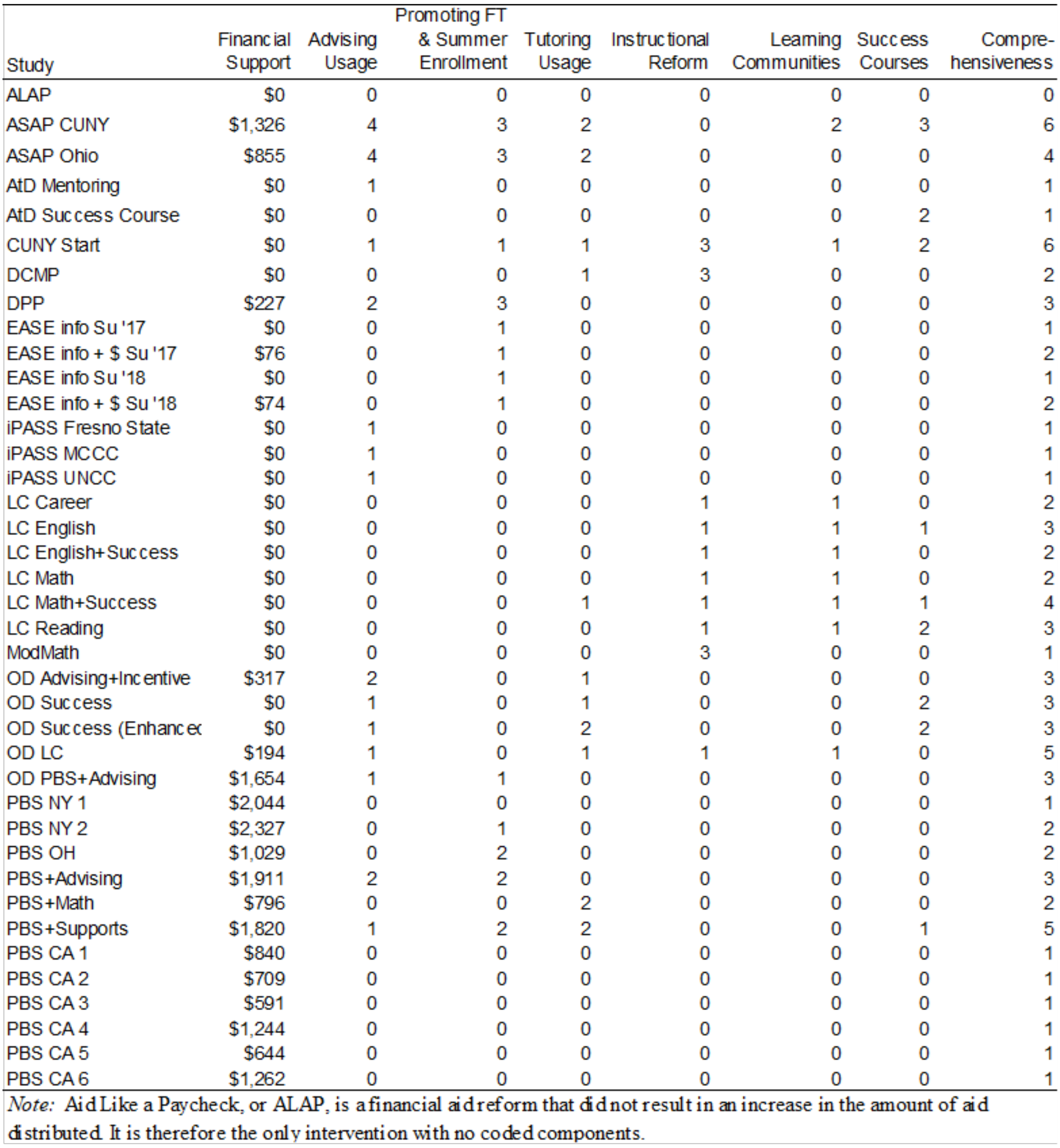
Intervention Component	Code Description
Increased Financial Supports	Cumulative amount received, through the end of year 1 (adjusted for inflation and regional pricing)
Increased Advising Usage	Cumulative additional advising visits, through the end of year 1:
	0 : Nothing
	1 : Very low + < 2 additional
	2 : Low + 2–7.9999
	3 : Med + 8–13.9999
	4 : High + 14 or more
Promoting Full-time & Summer Enrollment	# of terms intervention tries to influence enrollment intensity, through the end of year 1:
	0 : None
	1 : Require/incentive/encourage: FT Fall, FT Spring, OR Summer enrollment
	2 : 2 of the above
	3 : 3 of the above
Increased Tutoring Usage	Additional tutoring per semester, through the end of year 1:
	0 : None
	1 : 1 semester + < 3 additional visits, 1 semester none
	2 : 2 semesters + < 3 additional visits
	3 : 1 semester + ≥ 3 additional visits, 1 semester + < 3 additional visits
	4 : Average 3 additional visits or more in each semester
Instructional Reform	0 : None
	1 : 1 semester, inconsistent implementation
	2 : 1 semester, consistent implementation
	3 : (1 semester, multi-subject consistent) OR (1 semester consistent, 1 semester inconsistent)
Learning Communities	# of semesters of LCs offered (0, 1, or 2)

Description of Codes for Intervention Features (Cumulative Through Year 1) (*continued*)

Intervention Component	Code Description
Success Course	Additional participation per semester, through the end of year 1: 0 : None 1 : + 1—9.9pp took course 2 : + 10—49.9pp took course 3 : ≥ +50pp or higher
Comprehensiveness	Total # of intervention components that we studied, ranging from 0 to 6.

Appendix Table A3

Coded Values of Each Feature, by Intervention (Cumulative Through Year 1)



Study	Financial Support	Advising Usage	Promoting FT & Summer Enrollment	Tutoring Usage	Instructional Reform	Learning Communities	Success Courses	Compre-hensiveness
ALAP	$0	0	0	0	0	0	0	0
ASAP CUNY	$1,326	4	3	2	0	2	3	6
ASAP Ohio	$855	4	3	2	0	0	0	4
AtD Mentoring	$0	1	0	0	0	0	0	1
AtD Success Course	$0	0	0	0	0	0	2	1
CUNY Start	$0	1	1	1	3	1	2	6
DCMP	$0	0	0	1	3	0	0	2
DPP	$227	2	3	0	0	0	0	3
EASE info Su '17	$0	0	1	0	0	0	0	1
EASE info + $ Su '17	$76	0	1	0	0	0	0	2
EASE info Su '18	$0	0	1	0	0	0	0	1
EASE info + $ Su '18	$74	0	1	0	0	0	0	2
iPASS Fresno State	$0	1	0	0	0	0	0	1
iPASS MCCC	$0	1	0	0	0	0	0	1
iPASS UNCC	$0	1	0	0	0	0	0	1
LC Career	$0	0	0	0	1	1	0	2
LC English	$0	0	0	0	1	1	1	3
LC English+Success	$0	0	0	0	1	1	0	2
LC Math	$0	0	0	0	1	1	0	2
LC Math+Success	$0	0	0	1	1	1	1	4
LC Reading	$0	0	0	0	1	1	2	3
ModMath	$0	0	0	0	3	0	0	1
OD Advising+Incentive	$317	2	0	1	0	0	0	3
OD Success	$0	1	0	1	0	0	2	3
OD Success (Enhancec	$0	1	0	2	0	0	2	3
OD LC	$194	1	0	1	1	1	0	5
OD PBS+Advising	$1,654	1	1	0	0	0	0	3
PBS NY 1	$2,044	0	0	0	0	0	0	1
PBS NY 2	$2,327	0	1	0	0	0	0	2
PBS OH	$1,029	0	2	0	0	0	0	2
PBS+Advising	$1,911	2	2	0	0	0	0	3
PBS+Math	$796	0	0	2	0	0	0	2
PBS+Supports	$1,820	1	2	2	0	0	1	5
PBS CA 1	$840	0	0	0	0	0	0	1
PBS CA 2	$709	0	0	0	0	0	0	1
PBS CA 3	$591	0	0	0	0	0	0	1
PBS CA 4	$1,244	0	0	0	0	0	0	1
PBS CA 5	$644	0	0	0	0	0	0	1
PBS CA 6	$1,262	0	0	0	0	0	0	1

Note: Aid Like a Paycheck, or ALAP, is a financial aid reform that did not result in an increase in the amount of aid distributed. It is therefore the only intervention with no coded components.

Appendix Table A4

Estimated Distribution of True Average Effects, by Outcome and Semester

			Effect Distribution (across interventions)							
Outcome Measures	N	J	Mean (β)	SD (τ)	P10	P25	P50	P75	P90	γ
Enrollment (pp)										
Semester 1	65,604	39	2.8	3.5	-0.4	0.1	2.3	4.4	6.4	0.793
Semester 2	65,604	39	2.3	2.4	0.0	0.8	2.0	3.3	5.7	0.493
Semester 3	57,817	38	1.4	2.0	-0.8	0.6	1.4	2.1	3.9	0.346
Semester 4	54,509	38	1.2	1.8	-1.1	0.1	1.0	1.7	3.0	0.254
Semester 5	43,453	31	0.7	0.4	0.1	0.3	0.6	1.0	1.3	0.024
Semester 6	33,748	26	0.2	1.6	-2.4	-0.7	0.0	1.2	2.3	0.216
Credits Earned										
Semester 1	60,683	33	0.48	0.55	-0.03	0.13	0.48	0.60	1.29	0.790
Semester 2	60,683	33	0.43	0.57	-0.11	0.01	0.33	0.74	1.10	0.771
Semester 3	55,644	32	0.25	0.47	-0.17	-0.02	0.11	0.37	0.60	0.687
Semester 4	44,823	30	0.17	0.34	-0.14	-0.05	0.11	0.27	0.48	0.531
Semester 5	35,078	29	0.05	0.10	-0.08	-0.04	0.05	0.12	0.19	0.106
Semester 6	21,163	22	0.02	0.07	-0.05	-0.01	0.02	0.05	0.12	0.038
Cumulative # of Semesters Enrolled										
Semester 1	65,604	39	0.028	0.035	-0.004	0.001	0.023	0.044	0.064	0.792
Semester 2	65,604	39	0.052	0.049	0.003	0.019	0.036	0.088	0.119	0.660
Semester 3	57,817	38	0.064	0.068	0.000	0.015	0.048	0.097	0.168	0.579
Semester 4	54,509	38	0.077	0.082	-0.015	0.030	0.059	0.132	0.179	0.512
Semester 5	43,453	31	0.101	0.071	0.035	0.049	0.081	0.124	0.203	0.364
Semester 6	33,748	26	0.141	0.089	0.044	0.068	0.123	0.179	0.264	0.339
Cumulative Credits Earned										
Semester 1	60,683	33	0.47	0.54	-0.03	0.13	0.49	0.60	1.28	0.791
Semester 2	60,683	33	0.89	1.02	-0.09	0.23	0.73	1.25	2.02	0.801
Semester 3	55,644	32	1.14	1.52	-0.21	0.16	0.92	1.69	2.34	0.790
Semester 4	44,823	30	1.43	1.88	-0.21	0.25	1.15	2.07	2.70	0.759
Semester 5	35,078	29	1.70	1.93	0.30	0.64	1.11	2.22	2.92	0.643
Semester 6	21,163	22	2.46	2.20	0.34	1.51	2.20	2.67	4.90	0.542

Estimated Distribution of True Average Effects, by Outcome and Semester (*continued*)

			Effect Distribution (across interventions)							
Outcome Measures	N	J	Mean (β)	SD (τ)	P10	P25	P50	P75	P90	γ
Degree Earned (pp)										
Semester 4	40,323	23	0.9	2.4	-0.8	-0.5	0.3	1.0	3.4	0.716
Semester 5	36,782	23	1.6	4.1	-1.7	-0.7	0.3	2.4	5.4	0.809
Semester 6	30,613	20	1.7	5.9	-1.3	-0.6	0.3	3.1	9.7	0.863

Note. pp = percentage points; N = number of students; J = number of interventions; The mean (β) and standard deviation (τ) of the cross-intervention distribution of effects are estimated using the fixed intercept, random treatment coefficient (FIRC) model. Percentiles are based on adjusted empirical Bayes estimates of intervention effects (Bloom et al., 2017); γ is the ratio of the variance of the intervention specific empirical Bayes impact estimates to $\hat{\tau}^2$, as described in Appendix B.

Appendix B

Comparing a Current Impact Estimate to an Estimated Distribution of Past True Impacts

This appendix describes two scenarios for how to compare a current intervention impact estimate to an estimated distribution of past related true impacts. The first scenario involves a researcher who wants to compare a current impact estimate for a postsecondary intervention to an estimated distribution of corresponding true impacts for postsecondary interventions in Appendix Table A4 of this paper. The second scenario involves a researcher who wants to compare a current impact estimate for any type of intervention to a corresponding distribution of true impacts for related interventions that the researcher will estimate from past and current research findings. In both scenarios, the goal is to compare a current intervention impact to what is known about the distribution of true impacts for related interventions.

Scenario #1: Using an Estimated Impact Distribution From Appendix Table A4 as a Benchmark

Appendix Table A4 presents the mean, standard deviation, and percentile values for estimated distributions of true intervention impacts on postsecondary student outcomes. These findings represent distributions of adjusted empirical Bayes impact estimates ($\hat{\beta}_j^{AEB}$) across MDRC randomized trials of postsecondary interventions conducted during the past 20 years.

For example, based on data for previous MDRC trials of 33 postsecondary interventions, the estimated overall mean true impact ($\hat{\beta}$) on credits accumulated during students' first two semesters after random assignment is 0.89 credits, and the estimated standard deviation ($\hat{\tau}$) is 1.02 credits. Corresponding estimates of the 10th, 25th, 50th, 75th, and 90th percentile values for this distribution are –0.09, 0.23, 0.73, 1.25, and 2.02 credits, respectively.

To compare an OLS impact estimate ($\hat{\beta}_{j^*}^{OLS}$) from a current study (j^*) for the same student outcome to the preceding distribution, one must transform the OLS estimate to its adjusted empirical Bayes counterpart ($\hat{\beta}_{j^*}^{AEB}$). Conceptually, this transformation is a two-step process.[21] To be concrete, consider the process in the context of a current study where $\hat{\beta}_{j^*}^{OLS}$ equals 1.5 credits, and has an estimated standard error ($\widehat{se}(\hat{\beta}_{j^*}^{OLS})$ of 0.5 credit.

The first step is to convert $\hat{\beta}_{j^*}^{OLS}$ to a standard empirical Bayes impact estimate ($\hat{\beta}_{j^*}^{EB}$) using Equation B.1.

$$\hat{\beta}_{j^*}^{EB} = \hat{\lambda}_{j^*}\hat{\beta}_{j^*}^{OLS} + \left(1 - \hat{\lambda}_{j^*}\right)\hat{\beta} \tag{B.1}$$

where $\hat{\lambda}_{j^*}$ is the estimate of the reliability of $\hat{\beta}_{j^*}^{OLS}$, calculated as in Equation B.2.

$$\hat{\lambda}_{j^*} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + (\widehat{se}(\hat{\beta}_{j^*}^{OLS}))^2} \tag{B.2}$$

Thus, for the present example,

$$\hat{\lambda}_{j^*} = \frac{(1.02)^2}{(1.02)^2 + (0.5)^2} = 0.81$$

and

$$\hat{\beta}_{j^*}^{EB} = \hat{\lambda}_{j^*}\hat{\beta}_{j^*}^{OLS} + \left(1 - \hat{\lambda}_{j^*}\right)\hat{\beta} = 0.81\left(1.5\right) + 0.19\left(0.89\right) = 1.38$$

The next step is to convert $\hat{\beta}_{j^*}^{EB}$ to $\hat{\beta}_{j^*}^{AEB}$ as follows.[22]

$$\hat{\beta}_{j^*}^{AEB} = \hat{\beta} + \frac{1}{\sqrt{\gamma}}\left(\hat{\beta}_{j^*}^{EB} - \hat{\beta}\right) \tag{B.3}$$

where γ is an adjustment factor that accounts for the fact that the estimated variance of standard empirical Bayes impact estimates ($\widehat{var}(\hat{\beta}_j^{EB})$) for a sample of interventions systemically *understates* the corresponding variance of true impacts (τ^2). Therefore

$$\gamma \equiv \frac{\widehat{var}(\hat{\beta}_j^{EB})}{\hat{\tau}^2} \tag{B.4}$$

[21] Operationally, the two-step process can be represented by a single closed-form expression.

[22] See Bloom et al. (2016) for a discussion of adjusted empirical Bayes impact estimates.

where:

$$\widehat{var}\left(\hat{\beta}_j^{EB}\right) = \frac{\sum_{j=1}^{J}(\hat{\beta}_j^{EB} - \hat{\beta})^2}{J-1} \tag{B.5}$$

and J is the total number of intervention impact estimates in the benchmark sample.

Appendix Table A4 reports the value of γ for each postsecondary outcome measure represented. For credits accumulated during students' first two semesters after random assignment, γ equals 0.801.

Consequently, for the present example

$$\hat{\beta}_{j^*}^{AEB} = \hat{\beta} + \frac{1}{\sqrt{\gamma}}\left(\hat{\beta}_{j^*}^{EB} - \hat{\beta}\right) = 0.89 + \frac{1}{\sqrt{0.801}}\left(1.38 - 0.89\right) = 1.44 \text{ credits earned.}$$

This value lies between the 75th and 90th percentile values (1.25 and 2.02 credits, respectively) in Appendix Table A4. Hence, relative to past postsecondary interventions used to create the benchmark distribution for our example, the impact of the current hypothetical intervention is substantially positive.

A researcher could then interpolate the percentile value ($\hat{P}_{current}$) for the current intervention impact to determine where between the 75th and 90th percentiles it lies. For this purpose, a linear interpolation is probably a reasonable approximation. Thus, for the present example,

$$P_{current} = 75 + \left(\frac{1.44 - 1.25}{2.02 - 1.25}\right)\left(90 - 75\right) \approx 79$$

Hence, existing information indicates that the impact of the current intervention is comparable to the 79th percentile of the estimated distribution of previous related true impacts.

An online tool associated with this paper will make these calculations for a user—they simply need to input the effect estimate and its associated standard error from their study. See https://www.mdrc.org/the-rct-empirical-benchmarks.

Scenario #2: Estimating an Impact Distribution to Serve as a Benchmark

Now consider how to proceed when a researcher must estimate a benchmark distribution to help interpret a current intervention impact estimate. In this case, the researcher should include the current intervention in the benchmark distribution.[23]

To estimate this distribution, one needs the OLS impact estimate ($\hat{\beta}_j^{OLS}$) and its estimated standard error ($\widehat{se}(\hat{\beta}_j^{OLS})$) for the current intervention and each past intervention. With this information, it is possible to estimate the grand mean (β) and variance (τ^2) of true intervention impacts for the population represented by the analysis sample. To do so, one can estimate the following random-effects meta-regression using the "V-known" procedure in SAS, the Meta-reg procedure in STATA, or their equivalents in other software packages.[24]

$$\hat{\beta}_j^{OLS} = \beta + v_j + e_j \tag{B.6}$$

where:

v_j = intervention j's deviation from the overall mean true impact. This deviation is assumed to be independently and identically distributed across interventions, with a mean of zero and a variance of τ^2,

e_j = intervention j's estimation error, which is assumed to be independently and normally distributed across interventions, with a mean of zero and a variance of $(se\left(\hat{\beta}_j^{OLS}\right))^2$.

Existing software reports resulting estimates of the grand mean true impact ($\hat{\beta}$) and the standard deviation of true impacts ($\hat{\tau}$) across interventions plus an empirical Bayes impact estimate ($\hat{\beta}_j^{EB}$) for each intervention studied, including the current one.

This provides all but one parameter estimate needed to compare a current intervention impact estimate with an estimated distribution of past and current true impacts. The missing parameter estimate, $\hat{\gamma}$, can be computed from Equation B.4. However, to do

[23] Operationally, this approach simplifies the process because the empirical Bayes impact estimate for the current intervention is estimated automatically with those for past interventions. In addition, the approach makes conceptual sense, because it includes information about the present intervention with that for all past related interventions to estimate a distribution of true impacts for a relevant benchmark population. However, whether one includes a current intervention with past interventions is unlikely to affect one's results, unless: (a) there are very few past interventions, (b) the estimated impact of the current intervention differs markedly from those for past interventions, and (c) the precision of the estimated impact for the current intervention is much greater than that for past interventions.

[24] Equation B.6 is an aggregate version of the FIRC model used to estimate impact distributions in Appendix Table A4.

so, one must first compute $\widehat{var}\left(\hat{\beta}_j^{EB}\right)$ as follows from the empirical Bayes impact estimates reported for our meta-regression.

$$\widehat{var}\left(\hat{\beta}_j^{EB}\right) = \frac{\sum_{j=1}^{J}(\hat{\beta}_j^{EB} - \hat{\beta})^2}{J-1} \tag{B.7}$$

where J is the total number of empirical Bayes impact estimates involved.

The next steps are to:

- Compute an adjusted empirical Bayes impact estimate ($\hat{\beta}_j^{AEB}$) for each intervention (j) based on Equation B.3,
- Rank-order these adjusted empirical Bayes estimates,
- Convert each rank-ordered position to a percentile value, and
- Locate the adjusted empirical Bayes estimate for the current intervention and identify its percentile value.

For example, with 20 interventions in the distribution (19 past interventions plus the current one), the least positive adjusted empirical Bayes estimate would have a percentile value of 2.5 and the most positive estimate would have a percentile value of 97.5, with 5 percentage point increments for each intervention in between. Thus, if the adjusted empirical Bayes impact estimate for the current intervention had the 2nd most positive value, it would represent the 7.5th percentile of the distribution. The 3rd most positive estimate would represent the 12.5th percentile value, and so forth.