IES Institute of Education Sciences

Federal Efforts Towards Investing in Innovation in Education Through the i3 Fund: A Summary of Grantmaking and Evidence-Building

A Publication of the National Center for Education Evaluation at IES

U.S. Department of Education
Miguel Cardona
Secretary

Institute of Education Sciences
Mark Schneider
Director

National Center for Education Evaluation and Regional Assistance

Matthew Soldner   Marsha Silverberg
Commissioner   *Associate Commissioner*

Tracy Rimdzius
Project Officer

Federal Efforts Towards Investing in Innovation in Education Through the i3 Fund: A Summary of Grantmaking and Evidence-Building

February 2024

Barbara D. Goodson
Eleanor Harvill
Maureen Sarna
Kyla Brown
Abt Associates

Rachel McCormick
MC² Education LLC

Finding innovative educational strategies that work is important to help improve student learning and close equity gaps nationwide. The goal of the U.S. Department of Education's Investing in Innovation Fund (i3) was to build high-quality evidence about effective educational strategies and to expand implementation of these strategies. Between 2010 and 2016, i3 awarded 172 grants totaling $1.4 billion to universities, school districts, and private non-profit organizations. The i3 Fund intentionally awarded different types of grants either to develop and test new, innovative but as-yet-unproven strategies or to learn more about the circumstances under which previously tested strategies are effective. Grantees were required to fund independent evaluations that would meet high standards for quality. This report examines the quality and findings of the 148 i3 grants with completed evaluations and provides information about the focus of the strategies those grantees tested.

Key Findings

- ***Most i3 grantees tested new ideas and, across all grants, the most commonly tested strategies involved professional development for educators to improve instruction or reform schools.*** Over two-thirds of all i3 grants awarded were Development grants to test new promising strategies (115 grants; 67 percent). The remaining grants were Validation and Scale-up grants awarded to test previously proven strategies at a larger scale (57 grants; 33 percent). Over one-third of grantees (53 grants; 36 percent) described their objective in terms of improving the quality of the classroom curricula and instruction. The second most common objective described by i3 grants was turning around struggling schools (38 grants; 26 percent).

- ***Grantees largely met the i3 goal of producing rigorous evidence about whether their strategy worked.*** All but nine grants (139 grants; 94 percent) met the requirement for i3 evaluations to be independent, 93 percent (138 grants) met criteria used to assess the quality of implementation measurement, and 76 percent of the i3 evaluations (112 grants) met What Works Clearinghouse standards.

- ***However, few of the strategies that grantees implemented and subsequently evaluated improved student outcomes in grantee sites.*** While 68 percent of all i3 evaluations (101 grants) found that the educational strategy was implemented with adequate fidelity, only 26 percent of i3 evaluations found at least one statistically significant positive effect and no negative effects on students (39 grants).

Wide differences in student achievement between students from disadvantaged and more advantaged backgrounds have persisted over the last two decades.[1] To help close this gap, federal education policy has increasingly sought to expand the use of evidence-based strategies shown to improve important student outcomes. Two challenges to this goal have been the limited number of strategies that have rigorous evidence of effectiveness and a way to expand previously proven strategies to more students and schools.

Established under the American Recovery and Reinvestment Act of 2009, the Investing in Innovation Fund (i3) was designed to address these challenges with one of the first uses of a "tiered evidence" program in government.[2] The U.S. Department of Education (the Department) awarded three types (tiers) of grants that ranged from supporting the development and testing of new innovative strategies in a few places up through the testing of proven strategies on a larger, even national, scale. To document each tested strategy's implementation and effectiveness, the Department required all i3 grantees to fund independent rigorous evaluations.[3] This investment in i3 represented a landmark effort to expand the evidence base on what works in education.

The tiered grant structure of i3 aligned the amount of funding awarded to the quality of the existing evidence supporting the proposed strategies and the expected scale of implementation. At the lowest tier, *Development grants* were intended to be engines for innovation, encouraging grantees to propose new ideas without prior evidence of improving student outcomes and then test their efficacy on a local scale. *Validation grants* (middle tier) and *Scale-up grants* (highest tier) were expected to expand implementation of strategies supported by prior evidence and evaluate their effectiveness at a larger scale, with new types of students, and in new settings. Scale-up grants were also expected to document the feasibility of their strategies becoming a national model, considering the infrastructure changes and resources that this might require.

In 2010, the Department funded a study to investigate the nature of the strategies supported by i3 and to summarize the impact of the strategies on student outcomes. This final study report assesses the quality of and findings from the i3 evaluations to determine whether the nearly $1.5 billion investment in i3 resulted in the intended contributions to a rigorous evidence base on educational strategies.[4] Although i3 grantmaking ended in 2016, lessons learned can inform similar efforts under i3's successor, the Education Innovation and Research (EIR) program, which was created by Congress in the 2015 reauthorization of the Elementary and Secondary Education Act.[5] (Appendix Section A.1 provides more information on the i3 program and the grants awarded over the seven years of the program.)

Overview of the Study Design

Who was included?

The study sample includes 148 (86 percent) of the population of all 172 i3 grants awarded in seven annual cohorts (2010–2016).[6] The remaining 24 grants were not included in the study sample because their findings were not available in time for inclusion in this report.[7]

How was the study conducted?

Systematic reviews of the quality of the evidence produced by the i3 evaluations were conducted using the evidence standards set by the What Works Clearinghouse (WWC) overseen by the Department's Institute of Education Sciences (IES), as well as study-developed criteria for other quality factors.[8] The study also summarized what the i3 evaluations found about implementation and effectiveness of the strategies. See Appendix Section B.4 for more detail.

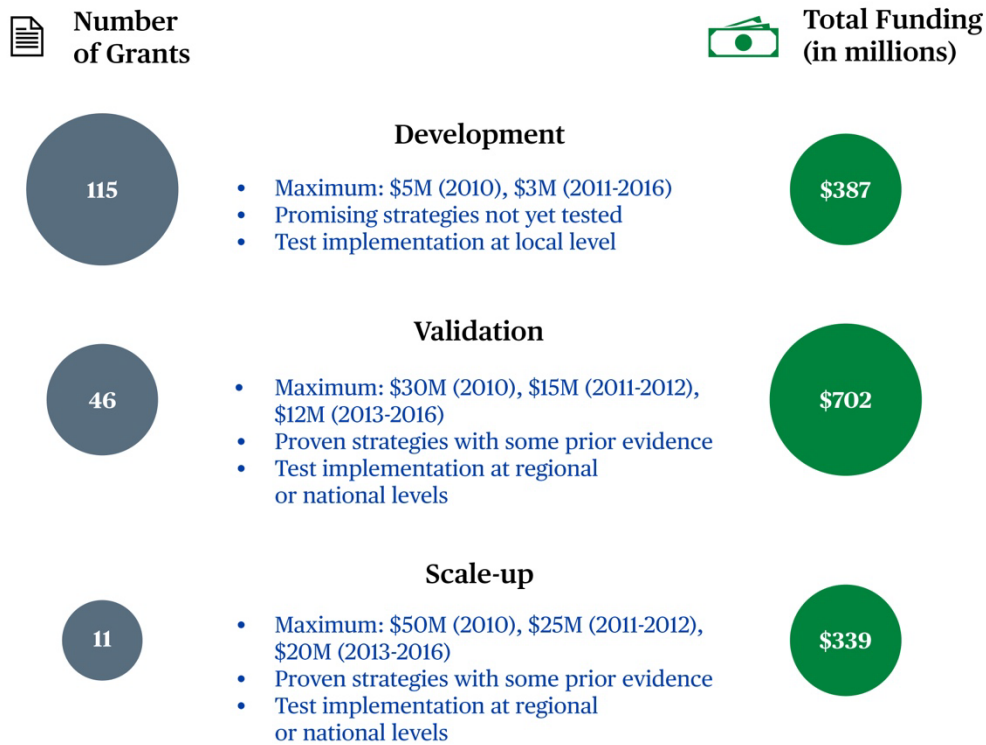**Which data were used and what was measured?[9]**

- Characteristics of grantees' strategies: Grantees provided graphical depictions of the intended relationship among the key components of their strategy (e.g., teacher professional development), their objective (e.g., strengthening instruction), and their desired student outcomes (e.g., greater reading proficiency). The study team collected grantee logic models and created a consistent set of codes to describe grantee activities, objectives, and intended outcomes.
- Independence of the evaluation: Each of the i3 grantees' independent evaluators reported on whether the grantee or strategy developer collected or analyzed outcome data for the evaluation.
- High-quality implementation data: The study team collected evaluation designs from i3 evaluators to determine whether evaluations met study-developed criteria for quality.
- Strength of evidence: The study team reviewed publicly available evaluation reports or data collected from i3 evaluators to determine an evidence rating for eligible study designs:[10]
  - Eligible designs:
    - *Randomized controlled trial*: Groups are created through a random process. If well implemented, this design results in groups that are similar on average across all characteristics, and any differences in outcomes between the groups are due to the educational strategy alone.
    - *Quasi-experimental design*: Groups are created through a non-random process. To be considered well implemented, these designs must demonstrate similarity between groups prior to implementation of the strategy.
  - WWC evidence ratings:
    - *Meets WWC Group Design Standards Without Reservations*: Indicates the highest degree of confidence that the strategy being evaluated caused the measured effects. This rating is reserved for findings based on well-implemented **randomized controlled trials**.
    - *Meets WWC Group Design Standards With Reservations:* Indicates a lower degree of confidence that the strategy being evaluated caused the measured effects. This rating applies to well-implemented **quasi-experimental designs** as well as randomized controlled trials that have problems with attrition.
    - *Does Not Meet WWC Group Design Standards*: Indicates a study not implemented rigorously enough to conclude with confidence that the strategy caused the measured changes in outcomes.
- Implementation fidelity: The study team reviewed publicly available evaluation reports and data collected from i3 evaluators to determine whether the grantees carried out their strategies as intended.
- Effect on student outcomes: The study team reviewed publicly available evaluation reports and data collected from i3 evaluators to determine whether the strategies had a positive effect on student outcomes.

MOST i3 GRANTEES TESTED NEW IDEAS AND, ACROSS ALL GRANTS, THE MOST COMMONLY TESTED STRATEGIES INVOLVED PROFESSIONAL DEVELOPMENT FOR EDUCATORS TO IMPROVE INSTRUCTION OR REFORM SCHOOLS.

To stimulate experimentation and build evidence in a wide variety of education topics, i3 allowed grantees broad choice in the strategies they proposed to implement and evaluate. For example, the proposed strategy could be a comprehensive program or model, a specific instructional practice, or a combination of practices. But the grant requirements and the Department's annual priorities[11] likely shaped which strategies were funded. For example, grantees were required to target their strategies to high-need students.[12] In addition, each year the Department specified a set of topics (called absolute priorities) that governed grantees' selection of strategies. The Department selected these absolute priorities to address persistent educational challenges and topics in need of more rigorous evidence. Consistent with the program's goal to build evidence about innovative strategies and take effective strategies to scale, the Department offered i3 grants at three tiers, as noted earlier. Examining the distribution of i3 grants–by tiers, objective, and key activities for improving student outcomes provides context for the findings of the i3 evaluations.

- *i3's commitment to testing strategies without prior evidence is seen in the large number of Development grant awards* (Exhibit 1). Across the seven cohorts of i3 grants, there were twice as many Development grants awarded to test new promising strategies (115 grants; 67 percent) as there were Validation (46 grants; 27 percent) and Scale-up (11 grants; 6 percent) grants to test previously proven strategies at a larger scale.

- *However, Validation and Scale-up grants received more of the i3 funds* (Exhibit 1). Because larger grant amounts were offered to grantees that were expanding implementation of strategies with prior evidence, the smaller number of Validation and Scale-up grants together received almost three-quarters of the total i3 funding despite making up only half of the total number of grants (57 grants; $1.041 billion out of $1.428 billion).

Exhibit 1: i3 Funding by Grant Type: 2010–2016

📄 **Number of Grants**

💵 **Total Funding (in millions)**

Development

115

- Maximum: $5M (2010), $3M (2011-2016)
- Promising strategies not yet tested
- Test implementation at local level

$387

Validation

46

- Maximum: $30M (2010), $15M (2011-2012), $12M (2013-2016)
- Proven strategies with some prior evidence
- Test implementation at regional or national levels

$702

Scale-up

11

- Maximum: $50M (2010), $25M (2011-2012), $20M (2013-2016)
- Proven strategies with some prior evidence
- Test implementation at regional or national levels

$339

Note: See Exhibits C.1a and C.1b in Appendix C for details. Data include all 172 grants. See Exhibits C.1c and C.1d in Appendix C for similar data for the 148 grants in the study sample.
Source: i3 Fund program documents

- ***The most common objective of the i3 grants was improving classroom curriculum and instruction*** (Exhibit 2). Over one-third of grantees (53 grants; 36 percent) described their objective in terms of improving the quality of the classroom curricula and instruction. The second most common objective among i3 grants focused on turning around struggling schools (38 grants; 26 percent).
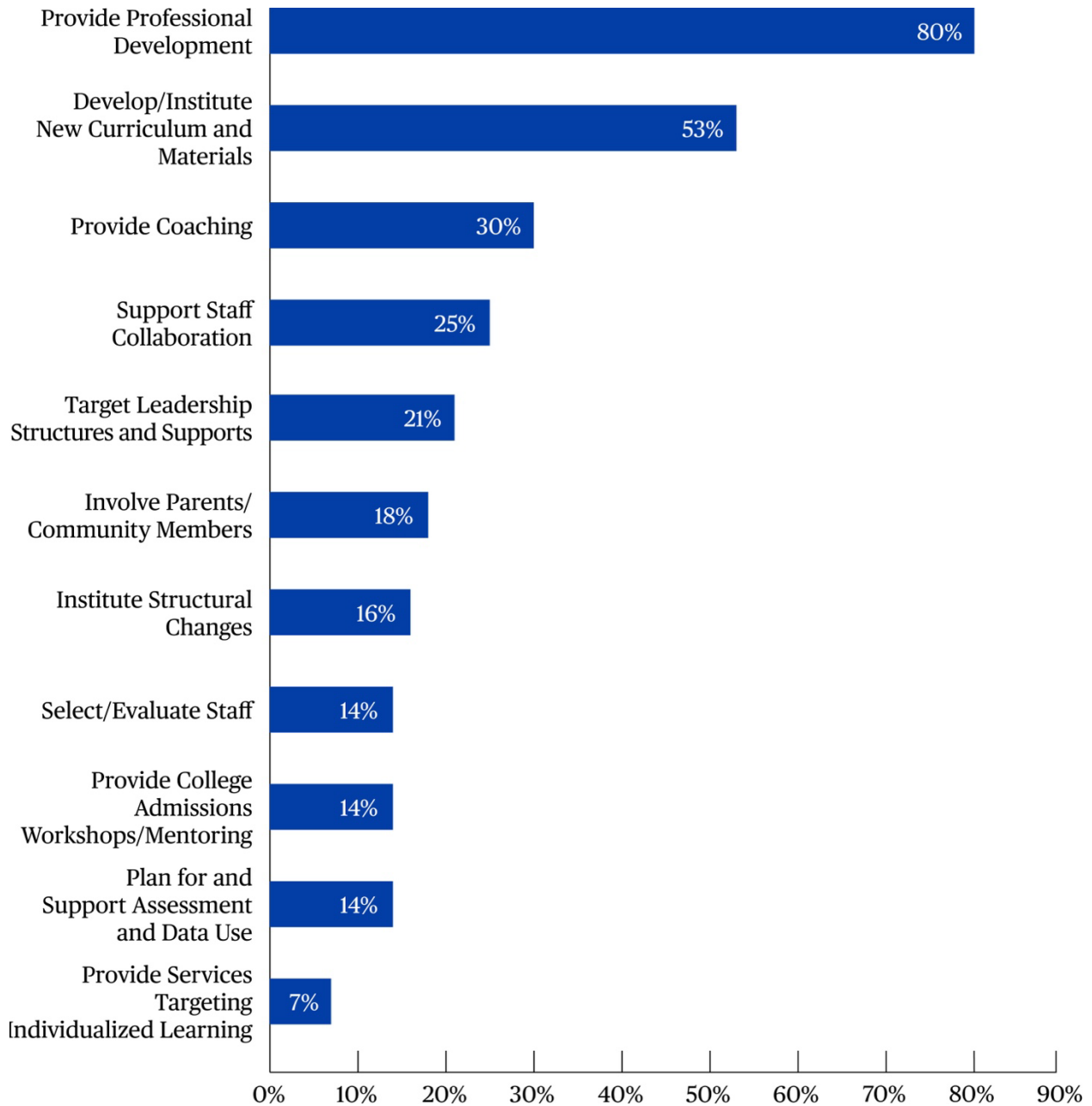
Exhibit 2: Objectives (Percentage of Grants)
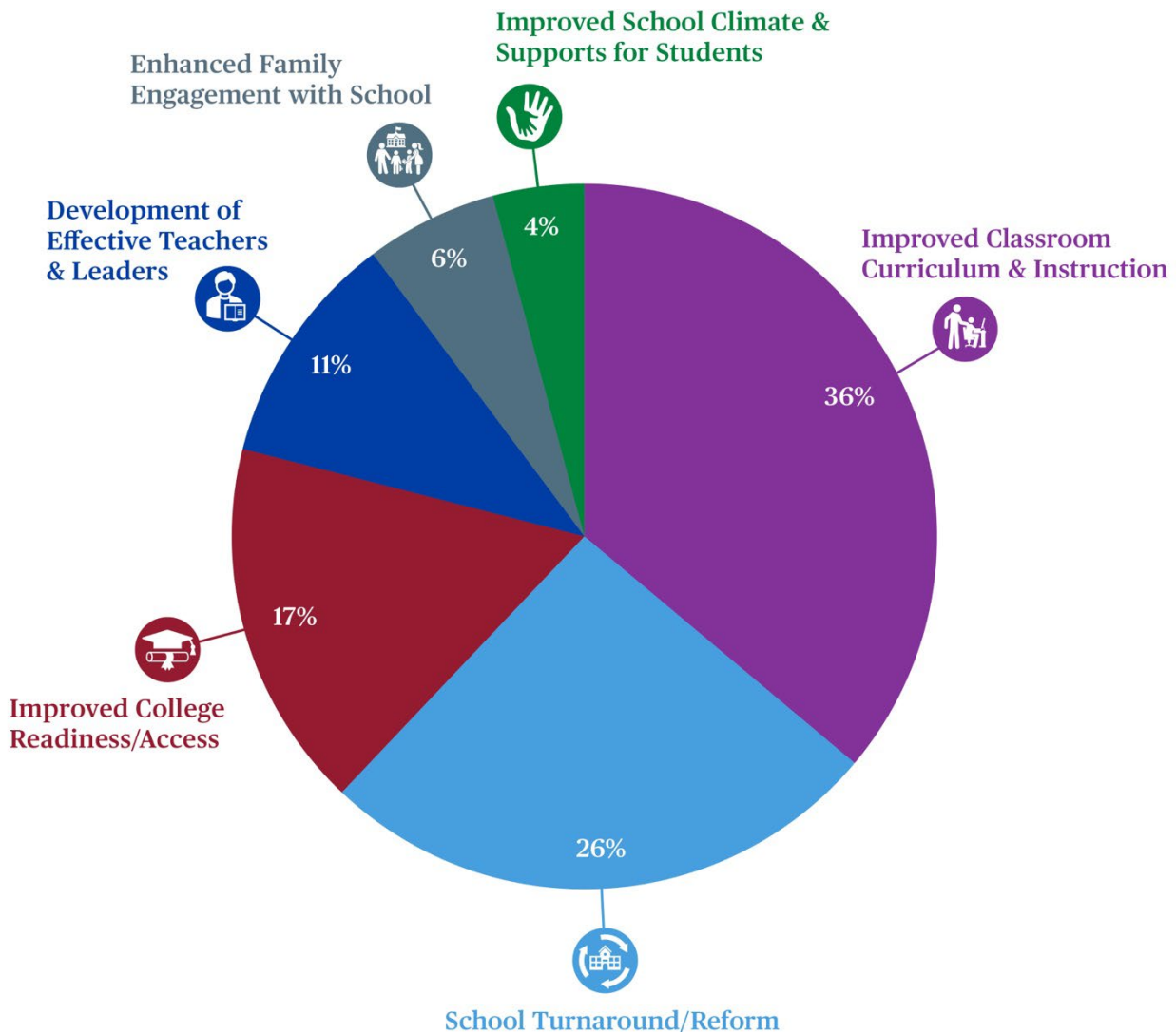
Note: See Exhibit C.2 in Appendix C for details. Data include objectives described by 148 grants in the study sample.
Source: Structured data collection from grants as part of implementation review

- ***Almost all i3 grantees delivered supports to teachers as a key component of their strategy*** (Exhibit 3). Eighty percent of grantees (119 grants) implemented some form of professional development to improve student success. The next two most commonly delivered supports were new classroom curricula (78 grants; 53 percent) and coaching for teachers (45 grants; 30 percent).

Exhibit 3: Key Components of Strategies (Percentage of Grants)

| Strategy	Percentage
Provide Professional Development	80%
Develop/Institute New Curriculum and Materials	53%
Provide Coaching	30%
Support Staff Collaboration	25%
Target Leadership Structures and Supports	21%
Involve Parents/Community Members	18%
Institute Structural Changes	16%
Select/Evaluate Staff	14%
Provide College Admissions Workshops/Mentoring	14%
Plan for and Support Assessment and Data Use	14%
Provide Services Targeting Individualized Learning	7%

Note: See Exhibit C.3 in Appendix C for details.
Data include key program activities proposed by 148 grants in the study sample.
Source: i3 grant evaluation reports

- *Reflecting program expectations, all i3 grantees examined effects on at least one student academic outcome* (Exhibit 4). The strategies proposed by i3 grantees were expected to improve student academic performance. Accordingly, all grantees targeted one academic outcome, and many grantees targeted multiple academic subject areas (71 grants; 48 percent). The single academic outcome targeted by the most grantees was English language arts (30 grants; 20 percent), followed by educational attainment (25 grants; 17 percent).

Exhibit 4: Targeted Student Academic Outcomes (Percentage of Grants)
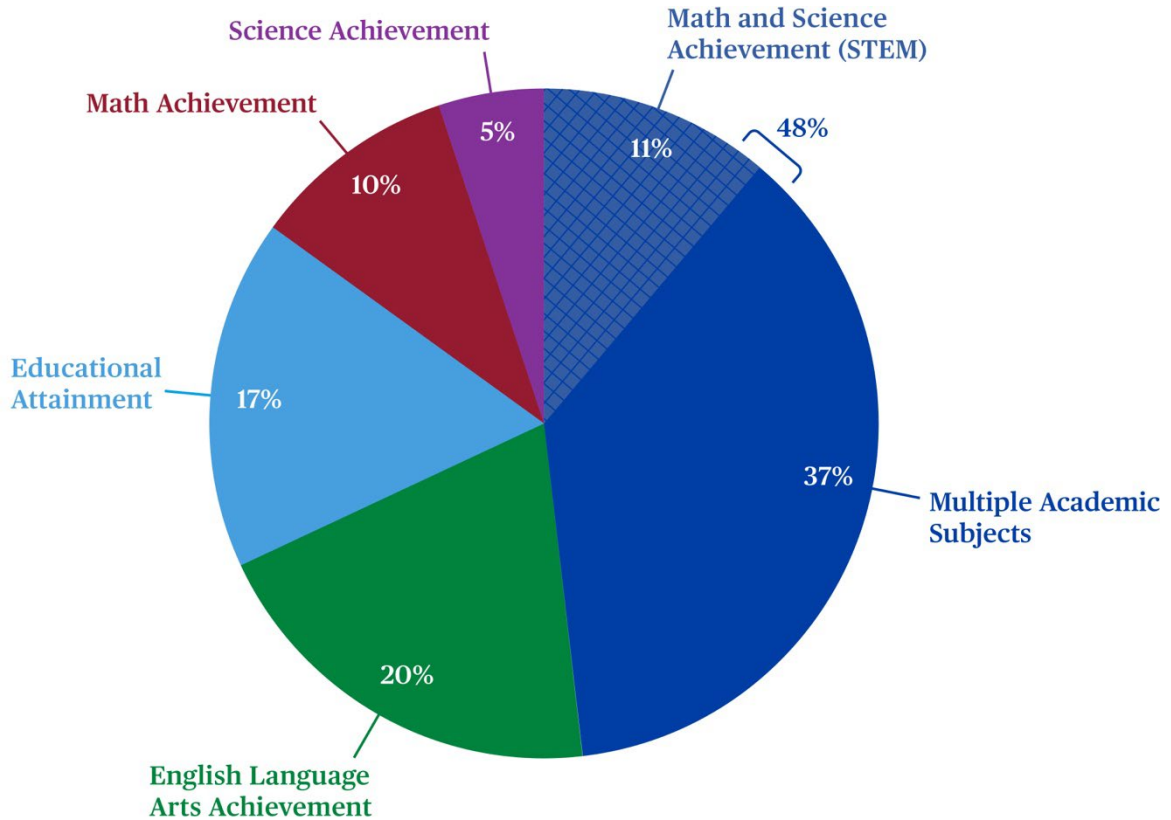
Note: See Exhibit C.4a in Appendix C for details. Exhibit includes outcomes targeted by the i3 Fund, including improved academic achievement, decreased dropout rates, or increased high school graduation rates.
Data include 148 grants in the study sample.
Source: i3 grant evaluation reports

Most grantees also targeted non-academic student outcomes such as student approaches to learning or engagement (112 grants; 76 percent) or teacher outcomes such as classroom practice or retention (87 grants; 59 percent). (See Exhibits C.4b and C.4c in Appendix C.)

GOAL OF BUILDING RIGOROUS EVIDENCE ABOUT IMPLEMENTATION AND EFFECTIVENESS OF i3 STRATEGIES LARGELY ACHIEVED

i3 aimed to significantly improve the amount and quality of information available to the field about whether strategies are effective, for which types of students, and in what contexts. Thus, the Department required all i3 grantees to fund independent evaluations to ensure the design and interpretation of results were objective. To maximize the rigor of the evaluations, the Department provided evaluation technical assistance (TA) to i3 evaluators and made participation in the TA a requirement of the grant. The Department also incentivized grantees' use of rigorous evaluation designs through the selection criteria for the federal grant competition. The Department expected Validation and Scale-up grants to fund evaluations designed to meet WWC standards so that the evidence of effectiveness generated from those grants could be used by other educators with some confidence. The Department's performance expectation for Development grants' evaluations was less rigorous,[13] but the evaluation TA provider helped strengthen the designs of the majority of the Development grantees that proposed designs in their grant applications that did not have the

potential to meet WWC standards (see Exhibit C.17 in Appendix C for details). Examining whether these expectations were met is important, given that the quality of evidence produced through i3 is critical to its success.

- ***The requirement for i3 evaluations to be independent was almost universally met.*** For nearly all i3 evaluations, the grantee staff implementing the i3-funded strategy and the developer were not involved in collecting or analyzing outcome data (139 grants; 94 percent) (Exhibit C.8 in Appendix C).

- ***Similarly, almost all i3 evaluations produced high-quality data on implementation.*** Ninety-three percent of i3 evaluations (138 grants; Exhibit C.6 in Appendix C) met all the criteria developed by the study team to assess the quality of implementation measurement (see Appendix Section B.4.2). Grantees were allowed flexibility in how to measure whether the key components of their strategy were implemented as intended. This measurement is important because knowing whether a strategy was put in place well helps in determining whether results from the evaluation are a "fair test" of the strategy.

- ***Most of the i3 evaluations met the Department's standards for rigorous evaluation*** (Exhibit 5). Seventy-six percent of the i3 evaluations (112 grants) met WWC standards. Twenty-five percent of the i3 evaluations (37 grants) were well-executed randomized trials that received the highest WWC rating (Meets Standards Without Reservations). Randomized controlled trials are considered the "gold standard" research method for determining the effectiveness of a strategy. Another 51 percent of evaluations (75 grants) met WWC standards with reservations (well-executed quasi-experimental designs or randomized trials that were not well executed[14]). The remaining 24 percent of i3 evaluations (36 grants) did not meet WWC standards. Nearly all of these were funded by Development grantees, for which rigorous designs were encouraged but not required.

Exhibit 5: Strength of Evidence from i3 Grant Evaluations, Overall and by Grant Type (Percentage of Grants)
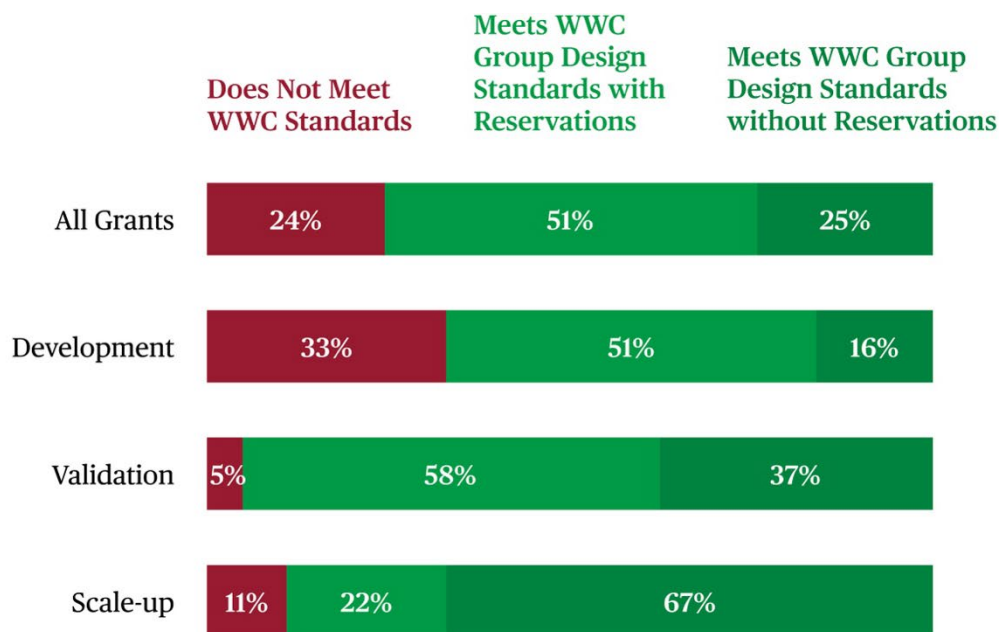
Note: Evaluation evidence ratings reflect the highest rating for any finding reported. These ratings are based on the version of the WWC Standards that was available at the time of the review. This figure presents data on all grants that received an official 3.0 or 4.0 rating from the WWC or an unofficial 3.0 or 4.0 rating from the i3 study team. Unofficial WWC version 3.0 reviews were replaced by more recent official WWC reviews when available. Unofficial ratings were based on the same standards and review procedures used by the WWC. See Exhibit B.12 in Appendix B for details.
Note: See Exhibit C.7 in Appendix C for details.
Data include 148 grants in the study sample.
Source: i3 grant evaluation reports

MOST GRANTEES DELIVERED THEIR STRATEGY AS INTENDED, BUT FEWER FOUND POSITIVE EFFECTS OF THEIR STRATEGY ON STUDENT OUTCOMES

Although rigorous evidence can help inform education practice even if positive effects are not found, i3 also hoped to identify new effective strategies and to replicate the positive findings of prior research on previously proven strategies. Grantees were required to make their evaluation findings broadly available[15] to improve the information available to other practitioners and policymakers about effective strategies. To assess whether i3 met its goal of expanding the evidence base on effective educational strategies, it is important to look not only at the quality of the evaluations but also at their findings.

- ***The majority of i3 grantees delivered their educational strategy as intended.*** Sixty-eight percent of all i3 evaluations (101 grants) found that the educational strategy was implemented with adequate fidelity. Excluding the seven percent of evaluations (10 grants) that did not meet the study's criteria for high-quality implementation data, the percentage of evaluations that found adequate implementation fidelity rises to 73 percent (101 grants; Exhibit C.10b in Appendix C). Most of the grantees that did not implement their proposed strategy as planned were Development grants (36 grants; 36 percent). This is not surprising given that many of these grantees were implementing their strategy for the first time in multiple sites and contexts.

- ***But a minority of i3 grantees' educational strategies yielded improved student outcomes*** (Exhibit 6). Twenty-six percent of all i3 evaluations found at least one statistically significant positive effect and no negative effects on students (39 grants).[16] More than three times as many Scale-up (5 grants; 56 percent) and Validation (19 grants; 48 percent) grantees as Development grantees (15 grants; 15 percent) found positive effects of their strategies–not surprising given that the strategies tested in these larger grants had prior evidence of effectiveness in other locales.

Exhibit 6: Effects of i3 Grant Strategies on Student Outcomes: i3 Grants Overall and by Grant Type (Percentage of Grants)
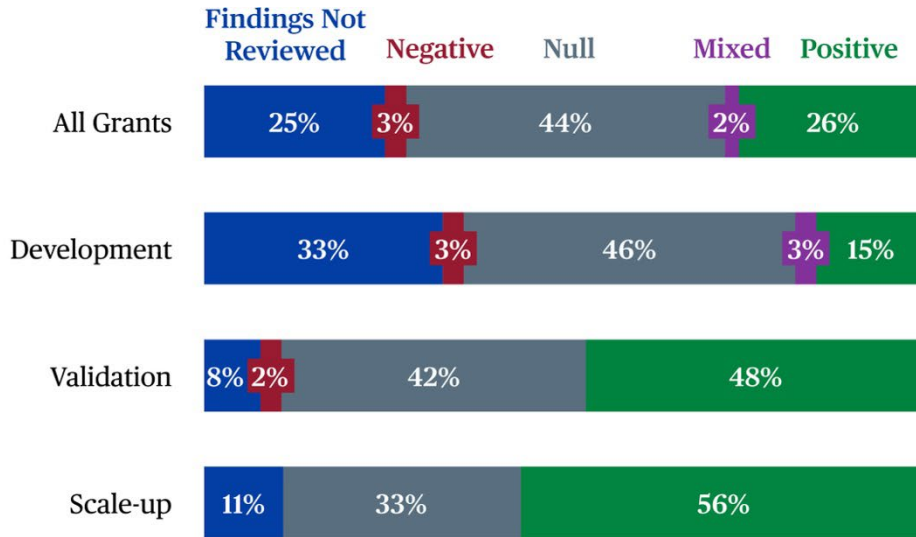
Note: This figure presents data on all evaluations that received an official 3.0 or 4.0 rating from the WWC or an unofficial 3.0 or 4.0 rating from the i3 study team. These findings are based on outcomes reviewed using the version of the WWC Standards and the associated review protocol that were available at the time of the review. The "findings not reviewed" category includes all grantees that did not meet WWC standards for at least one student outcome. One Validation grant that met standards without reservations is included in the "findings not reviewed" category because the WWC review focused exclusively on teacher outcomes and did not review student outcomes.

Positive–at least one positive finding and no negative findings after correcting for multiple comparisons. Negative– at least one negative finding and no positive findings after correcting for multiple comparisons. Mixed– at least one positive finding and one negative finding after correcting for multiple comparisons. Null–no statistically significant findings after correcting for multiple comparisons.

Note: See Exhibit C.11a in Appendix C for details.

Data include 148 grants in the study sample.

Source: i3 grant evaluation reports

- ***Almost half of the statistically significant positive effects on student performance were large*** (Exhibit 7). Across the most frequently assessed student outcome domains, 38 i3 evaluations found 48 statistically significant positive effects.[17] Nineteen of the 48 statistically significant positive effects on student performance (40 percent) had average effect sizes greater than 0.20 standard deviations (see textbox).[18] In addition to these large, statistically significant positive findings, ten i3 evaluations found 12 additional effects that were large but not statistically significant.[19]

Exhibit 7: Magnitude of Effects of i3 Grant Strategies on Student Performance, by Content Area
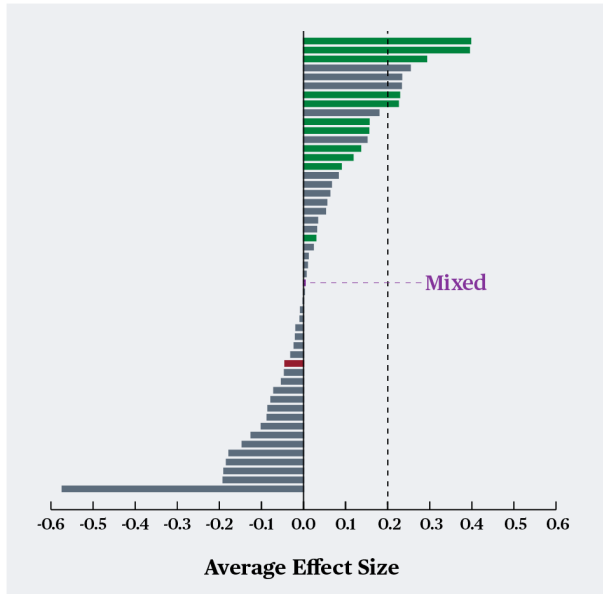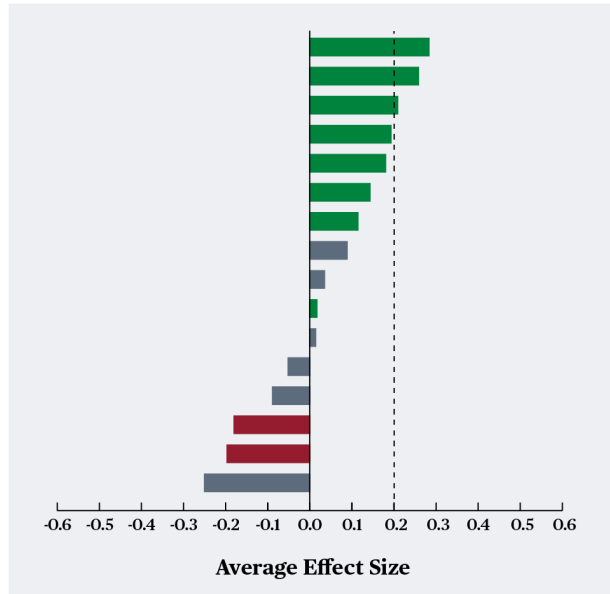
English language arts

Average Effect Size

Science

Average Effect Size

Mathematics

Mixed

Average Effect Size
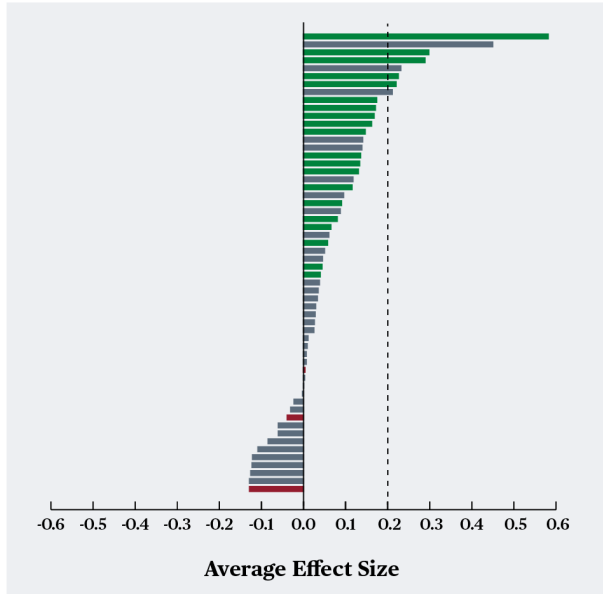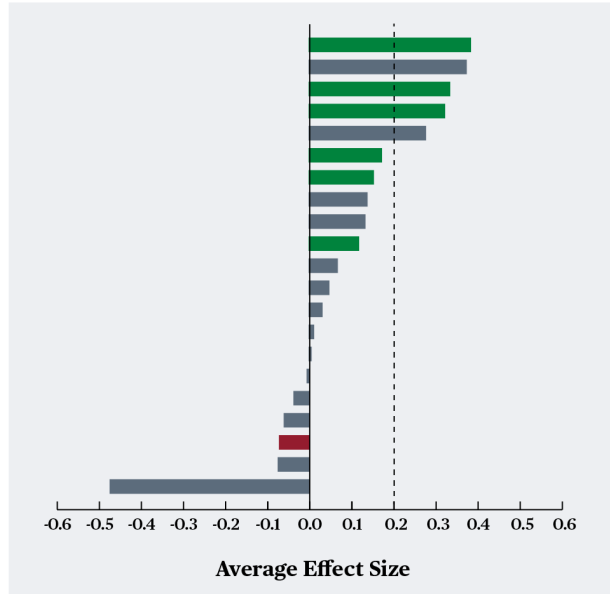
School attendance, progress, or attainment

Average Effect Size

Legend: ■ Positive ■ Negative ■ Null ■ Mixed

Note: Exhibit includes data on evaluations that met WWC standards for one or more student outcomes in each of four content areas according to an official 3.0 or 4.0 rating from the WWC or an unofficial 3.0 or 4.0 rating from the i3 study team. Each bar represents the effect of an educational strategy on the content area. For evaluations that targeted multiple outcomes in the same content area, the bar shows the average effect size across those outcomes. The exhibit omits one educational strategy targeting English language arts because the official WWC review did not report an effect size for this outcome and omits four educational strategies targeting school attendance, progress, and attainment because the average effect sizes were outside the range shown on the horizontal axis. Three of these omitted effect sizes were statistically significant and positive and the fourth was negative but not statistically significant. See Exhibits C.12a through C.12h in Appendix C for details.

Positive–at least one positive finding and no negative findings after correcting for multiple comparisons. Negative– at least one negative finding and no positive findings after correcting for multiple comparisons. Mixed– at least one positive finding and one negative finding after correcting for multiple comparisons. Null–no statistically significant findings after correcting for multiple comparisons.

Data include 102 grants with at least one outcome that met WWC standards in the content areas displayed: 59 evaluations measured English language arts, 51 mathematics, 21 science, and 20 school attendance, educational progress, or educational attainment. Source: i3 grant evaluation reports.

> **Effect Size in Research**
>
> *What is it?* In studies that compare groups offered new educational strategies with groups receiving the regular strategies normally provided, *effect size* is a quantitative measure of the size of the difference between the average outcomes in the two groups. Different researchers use different effect size formulas. Using the WWC definition, an effect size of 0.20 means that the difference in average outcomes between two groups is 20 percent of a standard deviation. Effect sizes allow readers to compare the size of differences on outcomes originally measured on different scales such as test score units and percentages.
>
> *How big is big?* The larger the effect size, the stronger the influence of the strategy being evaluated. One longstanding rule of thumb is to interpret effect sizes of 0.20 standard deviations or less as small, effects sizes of 0.50 as medium, and 0.80 as large.[20] However, smaller effect sizes may be educationally meaningful depending on the content area of the outcome, student age, when the outcome is measured, and the reliability of the outcome measure. A metric now being proposed for student gains in academic achievement is that effect sizes of less than 0.05 standard deviations is small, between 0.05 and 0.20 is medium, and greater than 0.20 is large.[21]

LOOKING AHEAD

The $1.4 billion awarded to i3 grantees was a significant investment in evidence-based grantmaking. Thus, it is important to understand the evidence produced by the i3 program. Reflecting on its successes and challenges can also provide lessons for the EIR program, the successor program to i3 that began awarding grants in 2017.

- *Did the emphasis on high-quality evaluation pay off?* Although the quality of the evaluations varied across the grant types, the vast majority of Scale-up and Validation grant evaluations met WWC standards. Over two-thirds of evaluations funded by Development grantees also met WWC standards, even though this was not explicitly encouraged by the Department until 2015. To put this achievement in context, evaluations funded by i3 grantees were almost twice as likely to meet WWC standards with or without reservations compared to all other studies the WWC has reviewed (76 versus 42 percent).[22]

  Beyond the WWC standards, the Department emphasized other factors that could make the i3 evaluations trustworthy and actionable. Three factors are notable. First, the independence requirement meant that developers and implementers, with vested interests in the effectiveness of their strategies, were not able to influence the findings. Second, the TA provider encouraged evaluators to pre-register their evaluation designs, and all but one evaluator complied. This helped to minimize the suppression of unfavorable findings (the "file drawer" problem) and post hoc revision to the analysis plan to find more favorable results ("fishing"). Lastly, the TA provider also helped evaluators meet the Department's expectation for high-quality implementation data, which involved clearly articulating the components of the strategies being tested and identifying a measurable threshold for adequate implementation. This emphasis on assessing fidelity of

implementation as part of effectiveness evaluations aligns with what the field increasingly sees as best practice in evaluation research.[23] Data on fidelity can help educators better understand the minimum level of implementation under which a strategy was found to have effects. These data can also offer important insights into potential improvements to the strategies and how to achieve fidelity of implementation in a broader range of school environments.

- ***Did the tiered structure incentivize evidence building?*** The tiered grant structure of i3 means that grantees whose evaluations met quality standards and found significant positive effects on student outcomes were eligible to apply for continued and increased funding in the next tier of grants. Thirty-eight percent of Development and Validation grantees whose strategies improved student outcomes successfully applied for and received grants in the next tier.[24] This encouraged an increase in the rigor of evaluations and the number of contexts in which the strategies were tested out.

- ***How can the program maximize learning from the grants designed to stimulate innovation?*** Although more of the i3 funds went to the top two tiers of grants, most of the grants funded each year of the program were in the lowest tier designed to stimulate innovation. These grants yielded fewer rigorous evaluations, hindering the ability to learn from these innovations. To improve the quality of the Development grantees' evaluations, the Department encouraged grantees to design studies that could meet WWC standards and also provided grantees with intensive evaluation technical assistance on how to do so. In 2015, the Department codified this encouragement by awarding more points in the grant competition to Development applicants who proposed rigorous evaluations in their applications. Accordingly, the proportion of Development evaluations meeting WWC standards increased over time.[25] For the successor EIR program, the Department explicitly encourages grantees in the lowest tier to use evaluation data to make changes to their strategies during the grant period. Encouraging grantees at all tiers to document how they adapt strategies to overcome implementation obstacles or tailor strategies to better meet student needs can yield important insights into how effective strategies can be taken to scale.

- ***How can evaluation technical assistance best support evidence-building?*** The Department invested $25 million over the life of i3 to provide grantees' independent evaluators with intensive TA on rigorous evaluation. Across all 172 grants, only one opted not to participate in the monthly calls with the TA team. The median number of calls per year between the evaluators and the evaluation TA team was nine, and all of the independent evaluators completed detailed implementation and impact study design plans, using tools provided by the TA team. In addition, some grantees and their evaluators increased the rigor of their initial evaluation designs during their participation in the TA. For example, 21 of the Development grantees proposed an evaluation design in their application that would not meet WWC standards. Working with the TA team on these designs, 14 of these 21 grantees developed revised designs that had the potential to meet WWC standards.[26] The Department did not formally evaluate the effectiveness of the TA.

Other programs considering the provision of evaluation TA may wish to formally measure the minimum intensity required to yield the kind of results achieved by i3.

- ***Was the strategy for synthesizing evidence across i3 strategies optimal?*** The strategies implemented and tested by i3 grantees almost always included combinations of key components or activities. These combinations of activities, as well as other characteristics such as their objectives, targeted grades, and outcomes, also varied extensively. To understand whether there might be a relationship between the characteristics of grantees' strategies and their effectiveness at improving student outcomes, the study team conducted an exploratory analysis (described in Section C.4.6 in Appendix C). This analysis found no clear relationship between any characteristics and whether the strategy produced positive findings.[27] This may have been due to the relatively small number of educational strategies i3 grantees tested compared to the number of different ways grantees bundled activities. This approach did not yield actionable evidence for the Department on whether and how it might either shape important features of future competitions or advise potential applicants in the design of their strategies. Future competitions that focus on a narrower set of outcomes, student populations, or strategies might improve the program's ability to synthesize learning across projects. Alternatively, or perhaps simultaneously, new approaches to synthesis may need to be developed to enhance the Department's ability to learn from a diverse set of activities.

- ***Did the larger i3 grants reach the targeted increase in scale?*** Validation and Scale-up grantees received larger awards to allow them to meet ambitious goals for expanding the use of effective practices, defined in terms of number of schools and/or students served. Validation grantees were expected to expand the reach of their strategies to the regional or national level, and Scale-up grantees to the national level. In addition, the grantees in these tiers were expected to identify and implement effective approaches to expansion, ideally while maintaining or increasing effectiveness and strong implementation across contexts. Notwithstanding some challenges recruiting participants at the scale proposed in their applications, most Scale-up grantees met their targeted number of students by the end of the grant. However, about two-thirds of the Scale-up grantees excluded a fairly large number of the schools, teachers, or students who received the strategy from the evaluation. Evaluators often did not explain how or why schools or individuals were excluded, which raises concerns about whether the evaluations measured impacts of the strategies as implemented at scale, as intended by the Department.[28] There was also little reporting on what was learned about effective approaches to scaling.

As education systems nationwide confront the long-term consequences of the COVID pandemic and address historical inequities in student achievement, innovative approaches to teaching, learning, and development are in high demand. Identifying innovations that work and taking them to scale is a core goal of tiered evidence programs such as the i3 Fund and its successor, EIR. Learning about and from these programs is critical if they are to achieve their full potential and yield the benefits students so desperately need.

ENDNOTES

1   S. Dynarski & K. Michelmore. (2017). Income Differences in Education: The Gap Within the Gap, *Econofact*. [Income Differences in Education: The Gap Within the Gap | Econofact](#).

2   See [GAO-16-818, TIERED EVIDENCE GRANTS: Opportunities Exist to Share Lessons from Early Implementation and Inform Future Federal Efforts](#) for information on other early tiered evidence programs in government.

3   *Independent* means that the evaluation is designed and carried out independent of, but in coordination with, any employees of the entities who develop a process, product, strategy, or practice and are implementing it. See Final Priorities, Requirements, Definitions, and Selection Criteria–Investing in Innovation Fund; Applications for New Awards; Investing in Innovation Fund, Development Grants; Rule and Notice, 78 Fed. Reg. 18704 (March 27, 2013).

4   A report summarizing the first 67 i3 evaluations to be completed, titled [The Investing in Innovation Fund: Summary of 67 Evaluations](#), was released in June 2018.

5   The EIR program–the successor to the i3 program–was created in 2015, as part of the Every Student Succeeds Act, which reauthorized the Elementary and Secondary Education Act. The EIR program retains the core purposes of i3 of development and testing of innovative strategies, as well as to test proven strategies on a larger scale, and retains the structure of awarding different types (tiers) of grants aligned to the quality of the existing evidence supporting the proposed strategies and the expected scale of implementation under EIR.

6   The 148 grants in the study sample represent 86 percent of all i3 grants funded and 90 percent of the total funds awarded by the program. The share of Development, Validation, and Scale-up grant types in the study sample did not differ significantly from their shares among all funded grants (*n*=172). The share of grants in the sample from later award cohorts was less than the share of grants from earlier cohorts. The share of grants in the study sample with evaluations that were designed to meet WWC standards, either with or without reservations, was lower, but did not differ significantly from the share of all i3 grants with such study designs. See Appendix Section B.2.2 for further details on how the study sample compares to all funded grants.

7   Only grantees that completed their evaluations and submitted their findings by August 2021 are included in the report. This date was selected to ensure sufficient time to complete What Works Clearinghouse reviews within the contract period for this study.

8   Different i3 evaluations were reviewed using different versions of the WWC standards and different review protocols. Six grants had multiple reviews in the WWC database. For them, the study selected the most complete, recent, and comprehensive review. A review was considered complete if all sections of the WWC study page (review details, findings, sample characteristics, and study details) were available. After prioritizing completeness, selecting the most recent review also happened to select the review that included the greatest number of findings, which made the selection process straightforward.

9   See Appendix Sections B.3 and B.4 for further details on the data sources and coding of the study measures.

10 For 44 of the 148 grantees included in this report, the study team used data submitted by i3 evaluators for the WWC evidence reviews. Though the evidence ratings for these evaluations are based on the same standards and review procedures used by the WWC, they are considered unofficial because the WWC reviews only publicly available reports.

11 The absolute priorities for each cohort of i3 grants are shown in Exhibit A.3 in Appendix A.

[12] The i3 program defined "high-need student" as a student at risk of educational failure or otherwise in need of special assistance and support, such as students who are living in poverty, who attend high-minority schools, who are far below grade level, who have left school before receiving a regular high school diploma, who are at risk of not graduating with a diploma on time, who are homeless, who are in foster care, who have been incarcerated, who have disabilities, or who are English learners.

[13] See the Department's performance measures for i3 at [2010-5139.pdf (govinfo.gov)](https://govinfo.gov).

[14] Of the Development grants that were rated as meeting WWC standards with reservations, 4 percent (6 grants) had designed evaluations with the potential to meet WWC standards *without* reservations but encountered high attrition at the cluster or individual level (1 percent), unallowed joiners (1 percent), or broken random assignment (1 percent). (For two of the Development grants with downgraded ratings, the reason was not clear in the WWC database.) Of the Validation and Scale-up grants that were rated as meeting WWC standards with reservations, 9 percent (14 grants) had designed evaluations with the potential to meet WWC standards *without* reservations but encountered high attrition at the cluster or individual level (4 percent), unallowed joiners (3 percent), or broken random assignment (2 percent), each of which meant the evaluations could not meet WWC standards without reservations.

[15] The Notice Inviting Applications (NIA) for the first three cohorts of i3 grants (FY 2010 through FY 2012) required grantees to "make broadly available through formal (e.g., peer-reviewed journals) or informal (e.g., newsletters) mechanisms, and in print or electronically, the results of any evaluations it conducts of its funded activities." The NIAs for the subsequent four cohorts of i3 grantees (FY 2013 through FY 2016) required grantees to make evaluation results broadly available "digitally and free of charge." See Section B.3.2 and Exhibit B.12 in Appendix B for details.

[16] When limiting the analysis to the 111 grants that met the Department's standards for rigor for at least one student outcome, a third (35 percent) found that their strategy improved at least one student outcome (Exhibit C.11b in Appendix C).

[17] Seven evaluations found statistically significant positive effects in more than one of the most frequently assessed student outcome domains (English language arts achievement, mathematics achievement, science achievement, and educational attainment). Five evaluations found statistically significant positive findings in two domains, one evaluation found statistically significant positive findings in three domains, and one found statistically significant positive findings in all four domains.

[18] Among the 19 large, positive statistically significant findings, five are in English Language Arts Achievement, five are in Math Achievement, three are in Science Achievement, and six are in Educational Attainment. These large, statistically significant positive findings represent 16 i3 evaluations.

[19] Among the 12 large effects that were not statistically significant, eight were positive (representing six i3 evaluations) and four were negative (representing four i3 evaluations).

[20] J. Cohen. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

[21] M.A. Kraft. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher, 49*(4), 241–253.

[22] As of April 2023, the WWC has published the reviews of 11,781 individual studies to its online database, accessible here: https://ies.ed.gov/ncee/wwc/ReviewedStudies#/. Of the 3,871 studies that received a WWC rating, 1,608 studies (42 percent) meet standards with or without reservations. The remaining 7,910 citations

were ineligible for review for a variety of reasons, such as because the publication serves as an additional source for information for a study review.

23 Institute of Education Sciences. (2021, June 10). Standards for Excellence in Educational Research (SEER). U.S. Department of Education. https://ies.ed.gov/seer/implementation.asp

24 This includes grantees that received a second (or third) higher tier award in either i3 or the successor program, EIR). See Exhibit C.16 in Appendix C for details. Note that there are other sources of additional funding for i3 grants. The National Center for Education Research funds grants studying similar topics as i3 and EIR do, including research that "contributes to improved education outcomes for all learners, and particularly for those whose education prospects are hindered by inadequate education services and conditions associated with poverty, race/ethnicity, limited English proficiency, disability, and family circumstance." https://ies.ed.gov/funding/overview.asp

25 See Exhibit C.17 in Appendix C for details.

26 See Exhibit C.18 in Appendix C for details.

27 The study team conducted a series of regression analyses to explore relationships between characteristics of educational strategies and their effectiveness at improving student outcomes. Each logistic regression produced a "likelihood ratio test" and an associated p-value for statistical significance, which indicates whether there is a systematic relationship between the set of characteristics included in the model and the outcome. If the likelihood ratio test is statistically significant, then this set of characteristics can help identify which educational strategies are more likely to have positive effects on student academic outcomes. Effects were not statistically significant and ranged from .10 to .39 (Exhibit C.19 in Appendix C).

28 See Exhibit C.9 in Appendix C for details.

ACKNOWLEDGEMENTS

DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST

The research team for this evaluation included staff from Abt Associates and its subcontractors, American Institutes for Research (AIR), ANALYTICA, Century Analytics, Mike Puma, LLC, Westat, and MC^2 Education LLC. None of the research team members has financial interests that could be affected by findings from this evaluation.