MATHEMATICA Policy Research

REPORT

Prospects for an Impact Evaluation of Project SEARCH: An Evaluability Assessment

June 10, 2016

Arif A. Mamun
Lori Timmins
David C. Stapleton

Submitted to:
National Institute on Disability, Independent Living, and Rehabilitation Research
Administration for Community Living
U.S. Department of Health and Human Services
330 C St SW, Room 1304
Washington, DC 20201
Project Officers: Hugh Berry
Contract Number: 90RT5034-02-00

Submitted by:
Mathematica Policy Research
1100 1st Street, NE, 12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Todd Honeycutt
Reference Number: 50209.00.500.032.000

This page has been left blank for double-sided copying.

ACKNOWLEDGMENTS

Funding for this study was provided by the Rehabilitation Research and Training Center on Vocational Rehabilitation Practices for Youth at TransCen, Inc., which is funded by the U.S. Department of Health and Human Services, Administration for Community Living, National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR) (Grant No: 90RT5034-02-00). The contents do not necessarily represent the policy of the U.S. Department of Health and Human Services and you should not assume endorsement by the federal government (Edgar, 75.620 (b)).

This report also benefited greatly from inputs from other colleagues at Mathematica. In particular, we would like to acknowledge feedback from Randall Brown, Tom Fraker, Todd Honeycutt, and David Wittenburg. We would also like to thank participants of Mathematica's Disability Affinity Group brownbag seminar. Stephanie Hulette provided excellent research assistance. Bill Garrett edited the report and Colleen Fitts formatted it.

This page has been left blank for double-sided copying.

CONTENTS

TABLES

This page has been left blank for double-sided copying.

EXECUTIVE SUMMARY

Project SEARCH has emerged as a promising program to address the challenges related to improving employment outcomes of youth with disabilities. It is a high school to work transition program that integrates employers and businesses with other educational and community rehabilitation service providers to engage youth with disabilities in paid work experiences. Recent monitoring and evaluation efforts suggest promising employment outcomes for Project SEARCH participants, but there has not yet been a rigorous impact evaluation with a large sample to demonstrate that these outcomes are substantially better than they would be if the participants had only relied on services and supports that are available outside of Project SEARCH.

In this report we present several design options for a rigorous impact evaluation of Project SEARCH. Relying on information we gathered from document reviews and from site visits conducted for this evaluability assessment, we propose two leading evaluation designs: one under the existing setting, where we take Project SEARCH sites, students, and other partners as given; and another under a demonstration setting, where we allow for the evaluation to play a role in determining the setting within which these players interact. We also discuss a few other alternative design options that we considered, but have concluded they are less appealing than those recommended.

- **Existing setting design.** Under the existing setting scenario, we propose a matched comparison group design with eligible youth from areas not served by Project SEARCH matched with individuals from areas served by the program.
- **Demonstration design.** Under a demonstration scenario, we propose a randomized experimental evaluation with school districts/local education agencies (LEAs) randomly assigning youth enrolled in the demonstration either to a treatment group that would have the opportunity to apply for Project SEARCH services, or to a control group that would have the opportunity to receive usual services from the state vocational rehabilitation (VR) agency.

For practical reasons, we recommend pursuit of the existing setting design first. We believe this design would meet the standards of rigor necessary for the findings to be credibly used to inform policy and, importantly, would be by far the most feasible and least expensive to implement. In particular, it would use existing Project SEARCH sites and participants for the intervention group, and would also use existing data sources (state VR agency, the Social Security Administration, the American Community Survey, U.S. Department of Education's Common Core of Data). Not only would these design features contain the costs of an evaluation, but they would also minimize some of the uncertainties and upfront coordination necessary to implement a design in the demonstration setting. Nonetheless, this design would still require careful planning and substantial effort. Specifically, we must determine exactly which Project SEARCH sites to focus on and the years of study; obtain research agreements with state VR agencies to access and link their data to SSA data; and gather data from other data sources to identify a credible comparison group.

Even though the demonstration setting design uses what program evaluators would call the gold standard for impact evaluation, a randomized controlled trial (RCT), it also would involve significantly more resources and a longer time frame to implement. In this design, school

districts would randomly assign youth either to a treatment group that would have the opportunity to apply for Project SEARCH services or to a control group that would have the opportunity to receive usual VR services. As such, we would need strong collaboration and buy-in from the local school districts and must obtain informed consent from each participant. Further, this design would face more stringent IRB approval requirements, and we would be required to collect and store at least some data from participants. We would likely be unable to obtain sample sizes as large as the existing setting design, making it more difficult to detect smaller impacts. These steps all involve significant resources and time to implement, leading to some uncertainty about its feasibility and success.

As in all social program evaluations, there are various threats under both the existing and demonstration setting design that could undermine the evaluation's ability to draw meaningful conclusions for policymakers and other stakeholders. However, the possibility of these threats materializing is not unusually or exceptionally high, and not apparently larger for one design than the other. Given that there is some risk in pursuing either of these designs helps tip the scale in favor of the design that would use far less resources, the existing setting design.

Implementing either of the leading evaluation designs would require collaboration with Project SEARCH and other entities. If an impact evaluation of Project SEARCH is pursued, we envision that it would require further discussion with the Project SEARCH leadership team as well as site staff. In addition, researchers involved in the evaluation would need buy-in from partners in each Project SEARCH site. The research team would also need to establish research and data use agreements with the participating state VR agencies and with SSA.

I. INTRODUCTION

Project SEARCH has emerged as a promising program to address the challenges related to improving employment outcomes of youth with disabilities. It is a high school transition program that integrates employers and businesses with other educational and community rehabilitation service providers to engage youth with disabilities in paid work experiences. Even though recent monitoring and evaluation efforts suggest promising employment outcomes for Project SEARCH participants, there has not been any impact evaluation of the program to date that is based on a large sample and meets the highest standards of rigor.

Supporting the successful transition of young people with disabilities into the world of work and to a path of economic self-sufficiency has been of keen policy interest in recent years. As youth with disabilities face special challenges beyond the issues facing all transition-age youth, the value of early work experience has been long recognized in research and practice in the field of special education and transition (D'Alonzo 1978; Halpern 1985; Carter, Austin, and Trainor 2012; Madaus et al. 2013; Wehman et al. 2015). Providing transition-age youth with disabilities with work experience during adolescence is a core component of current transition frameworks (National Alliance for Secondary Education and Transition 2005; National Collaborative on Workforce and Disability for Youth 2009). It has also emerged as a key topic in national policy initiatives; for example, the federal government has undertaken two initiatives with a sharp focus on providing youth with paid work experience—the Youth Transition Demonstration, and Promoting Readiness of Minors in Supplemental Security Income (PROMISE) (Fraker et al. 2014a, b).

However, actual employment outcomes for youth with disabilities have not improved over the years, and rigorous evidence on successful strategies that produce sustained improvements in youth outcomes is limited. Research suggest that the population of youth with disabilities have been plagued by unemployment and underemployment for decades (Butterworth et al. 2014). For example, data from the American Community Survey (ACS) suggest that in 2013, only 23 percent of young people with disabilities ages 16 to 21 and 41 percent of adults with disabilities ages 22 to 30 were employed (Butterworth and Migliore 2015). Numerous descriptive studies point to factors associated with better youth employment outcomes (Test et al. 2009). A number of smaller-scale intervention evaluations have identified effective avenues for increasing youth employment (Balcazar et al. 2012; Carter et al. 2009, 2011; Wehman et al. 2014). Rigorous studies of the Youth Transition Demonstration found that providing work-based experiences and system linkages, along with promoting youth empowerment and family involvement, can significantly improve short-term employment outcomes (Fraker et al. 2014b; Hemmeter 2014).

It is in this context that the current report presents ideas for a rigorous evaluation of Project SEARCH. We assess the feasibility of rigorous evaluation designs that would allow estimation of the impact of Project SEARCH by comparing the outcomes of youth served by the program with outcomes for a comparable group—a group that shows what the Project SEARCH participants' outcomes would have been in the absence of the program. For the evaluability assessment, we conducted two site visits with Project SEARCH staff and partners—one in Cincinnati, Ohio and another in Orlando, Florida—in December 2015 and April 2016, respectively. We also reviewed previous research related to Project SEARCH and other program documents. Although some evidence exists from a small randomized controlled trial involving

44 youth with autism spectrum disorder (Wehman et al. 2014) and from other descriptive studies (Christensen et al. 2015; Müller and VanGilder 2014), additional rigorous evidence would give policy makers more confidence in the program's efficacy. A study based on a larger number of youth with a range of disabling conditions and from a larger geographic area would shed light on the efficacy of Project SEARCH for a group of individuals who are more representative of current program participants. We assess the possibility of identifying a valid comparison group required for a rigorous impact evaluation, and availability of data for the youth who participate in Project SEARCH services as well as those in the comparison group.

A. Core components of Project SEARCH

Project SEARCH is an intensive job training program for high school students with disabilities. The history and detailed overview of the program is provided by Daston, Riehle, and Rutkowski (2012). The Project SEARCH program model was developed at the Cincinnati Children's Hospital Medical Center in 1996, and has expanded to 269 sites across 34 states in the U.S. and in four other countries in 2014. A Project SEARCH site typically accommodates approximately 12 students in each cohort; although the number may vary between 6 and 15, depending on the minimum number of students required to cover the expense for staffing.

As a business-led intervention, Project SEARCH aims to provide real-life integrated work experience to the participating students while meeting a real business need of an employer. It began with the simple intention of hiring people with disabilities. The issues of fairness and opportunities for youth with disabilities entered into the design of the program, but the primary motivation was to improve productivity and retention in high-turnover support positions.

Daston, Riehle, and Rutkowski (2012) note 10 core elements of the Project SEARCH model: (1) identifies key program outcome as integrated employment for each participant; (2) involves true collaboration among agencies supporting youth with disabilities; (3) is led by a business; (4) provides consistent on-site staff in the business place by partners; (5) serves young adults with significant disabilities; (6) relies on braided funding streams through redirected existing funds from different sources to make the program sustainable; (7) requires total immersion of participants in the work place; (8) collects outcomes data and records it in a national Project SEARCH database; (9) provides effective follow-along services to participants to retain employment; and (10) operates each site under a licensing agreement with Project SEARCH Cincinnati.

Students participating in Project SEARCH are embedded in a large community business, receiving real-life work experience and in-classroom instructions on employability skills. Students rotate through three 10-12 week unpaid internships within the business over one school-year, accruing approximately 720 hours of internship time and 180 hours of classroom time at the business learning competitive, marketable, and transferrable skills (Schall 2013). During the second half of the school year, participants receive job coaching and job development support that help refine skills, and carry out individualized job search. Participants are on site at the business each day for a minimum of six hours. Typical host businesses include large hospitals, hotels and resorts, municipal departments, and banking centers.

Project SEARCH is a collaborative model, requiring the involvement of multiple community partners to support youth in obtaining job training experiences. Strong collaboration must occur between students and their families, the local education agency (LEA) such as school districts, the state vocational rehabilitation (VR) program, a local community rehabilitation program (CRP), and a host business. Each partner plays a specific role. Students and their family members identify their personalized employment goals and participate in vocational assessments. The host business provides internship sites and a classroom for instruction. The LEA's role is to provide a teacher to implement the senior year Individualized Education Program (IEP) of student interns in the program and to provide classroom instruction of the employability skills curriculum. The state VR agency provides funding and supervision for job coaching services that are provided to students throughout the internships. Finally, the CRP provides the job coaches to assess student interests, provide on-site job coaching during the school day, and develop and supervise internships.

The cornerstone of Project SEARCH is total workplace immersion. The model involves an extensive period of training and career exploration, long-term job coaching, and continuous feedback from teachers, job coaches, and employers. At the completion of the training program, many students are employed in complex and rewarding jobs, the types of jobs in which very few workers are youth with disabilities.

B. Target population and recruitment strategy

Typically, the students that participate in Project SEARCH are those with intellectual and developmental disabilities who are in their last year of high school and transitioning to adult life. In addition to students with significant intellectual and developmental disabilities, the program serves students with a variety of other disabling conditions acquired before age 22 (for example, visual impairment, hearing impairment, orthopedic impairment, autism). The students fall in the age range of 18 to 22, usually receive special education services in school through an IEP, and have completed all of their high school credits and graduation or certification requirements, but are in deferred graduation status under the Individuals with Disabilities Education Act (IDEA). Young adults who aged out of high school can fill vacant slots in a Project SEARCH program if enough eligible high school-aged students cannot be identified. A key eligibility criteria is a personal and family interest in achieving competitive employment for the youth.

More specifically, following are the eligibility criteria for participating in Project SEARCH outlined in Daston, Riehle, and Rutkowski (2012):

- Participant is at least 18 years old
- Has completed high school credits necessary for graduation/certificate
- Agrees that this will be the last year of student services and will accept diploma/certificate at the end of Project SEARCH
- Meets eligibility requirements for VR
- Meets eligibility requirements for developmental disabilities services and other service providers as necessary for follow-along services (preferred but not necessary)
- Has independent personal hygiene and grooming skills

- Has independent daily living skills
- Is able to maintain appropriate behavior and social skills in the workplace
- Is able to take direction from supervisors and modify performance or change behavior, as requested
- Is able to communicate effectively
- Can utilize public transportation when available, and participate in travel training to maximize independence in travel
- Has previous experience in a work environment (including school, volunteer, or paid work)
- Is able to pass drug screen and felony checks and have immunizations up to date
- Desires and plans to work competitively in the community at the conclusion of the Project SEARCH program

Project SEARCH is perceived to work best when it is offered as part of a continuum of transition services (Daston, Riehle, and Rutkowski 2012). In general, students who succeed in the program are those who join the program after spending one or two years in more traditional career and technical education programs that allow for maturation, functional skill development, and career exploration.

As part of its recruitment effort, the program conducts information sessions for prospective participants, families, and school special education staff. Given that Project SEARCH is considered most effective at a later stage of a continuum of transition services, it is not surprising that students are typically referred to the program by their schools. VR counselors and other service providers may also be the source of a referral. Sometimes parents contact the program directly after learning about the program.

The selection process involves students applying in the winter and spring in the year preceding the program and application review by a committee that represents all partners and the host business. After initial review of the applications, eligible applicants are invited to tour the program and participate in hands-on assessments. The applicants are then interviewed and scored by the selection committee using an eligibility rubric. The eligibility rubric has 17 components or strands covering various dimensions of the eligibility criteria. The student is scored on a scale of 1 to 5 for each component, such that the highest possible score is 85. Many Project SEARCH sites look for candidates who score in the 50 to 70 range. Students above the range may not require Project SEARCH services, whereas those below the range may not have the functioning and employability skills necessary to be able to take advantage of the program services. However, this rule is not used consistently across sites, as local conditions may lead to alternative adaptations of the rubric (through additions to or subtractions from the standard rubric template), or the threshold score might be applied flexibly depending on the number of open slots and number of applicants in a particular year. At the end, the rubric is one of several ways the selection committee evaluates the applicant's potential for success. After the selection committee reviews the rubric score, the process for determining VR eligibility begins, which is typically completed over the summer so that VR eligibility is determined before the student begins the program.

Considering the factors described above that go in to the selection of students who participate in Project SEARCH, the participants are likely to be students with disabilities who would be considered relatively high-functioning. That also appeared to be the case when we observed participants in three Project SEARCH sites during the site visits we conducted as part of the current evaluability assessment.

C. Expected outcomes

Project SEARCH defines a successful outcome as competitive employment in the community with at least 20 hours per week and pay at the wage rate prevailing in the local labor market. With Project SEARCH participants experiencing total immersion in the workplace, the program aims to prepare students to exit high school to enter competitive, integrated employment. The model involves an extensive period of training and career exploration, long-term job coaching, and continuous feedback from teachers, job coaches, and employers. As a result, at the completion of the training program, many students are employed in competitive paid jobs. Recent Project SEARCH monitoring data reflect the program goal of integrated employment for each participant—in school-year 2013-14, 67 percent of the participants engaged in paid employment after completing the program. In addition, a recent randomized controlled study on the effectiveness of Project SEARCH involving 44 youth with autism spectrum disorder found that 21 of 24 youth (87 percent) who received Project SEARCH services and autism supports achieved employment, whereas only 1 of 16 youth (6 percent) in the control group achieved employment one year after completion (Wehman et al. 2014).

This page has been left blank for double-sided copying.

II. A RIGOROUS IMPACT EVALUATION DESIGN: SETTING THE STAGE

A critical challenge for conducting a rigorous impact evaluation of Project SEARCH would be to identify two groups of youth who are similar in all respects except for Project SEARCH participation. Students who participate in the program are a select group who are unlikely to be similar to the nonparticipants because students with disabilities who apply to the program are a self-selected group and because the program uses a flexible but detailed set of eligibility criteria to determine which applicants actually receive Project SEARCH services. In addition, school districts and areas where the program is offered may have different educational and labor market opportunities than those that do not. Consequently, simply comparing outcomes for students who participate in the program versus the nonparticipants is unlikely to produce an unbiased estimate of program impacts. This is the key challenge we try to address in identifying alternative evaluation designs for estimating the impacts of Project SEARCH on the students' employment related outcomes.

We present evaluation designs for estimating impacts of Project SEARCH under two different settings: (a) the existing setting, where we take the Project SEARCH sites, students, and other partners as given; and (b) a demonstration setting, where we allow for the evaluation to play a role in determining the setting within which these players interact to allow for a more rigorous evaluation of Project SEARCH. In each setting, we have identified a leading evaluation design, which is the most robust and rigorous design we consider to be feasible. The existing setting design involves a quasi-experimental approach with a matched comparison group constructed from potential comparison students in communities where Project SEARCH isn't offered; the matching approach used would take into account characteristics of the community as well as of students who are likely to participate in the program. The demonstration setting design involves a randomized controlled trial (RCT) approach in which individual students within a school district would be randomly assigned either to a treatment group with an opportunity to apply to Project SEARCH, or to a control group with an opportunity to receive usual services provided by the state VR agency. We also discuss a few other potential evaluation designs we have considered, but found less attractive. Below we discuss the key factors we have taken into account in developing the evaluation designs, followed by a discussion of the leading evaluation designs under each setting as well as other potential designs we have considered.

A. Key considerations for impact evaluation design

We considered several factors in determining alternative rigorous impact evaluation strategies. These include: (1) barriers to implementation, (2) expectations about program outcomes, (3) identifying a credible comparison group, (4) availability of data on an evaluation sample before engagement with the program, (5) availability of outcomes data on an evaluation sample, and (6) statistical power to detect meaningful impacts.

We conclude that the two leading evaluation designs we propose would meet the key evaluation considerations related to barriers to implementation, expectations about program outcomes, and identifying a credible comparison group. There are no major barriers to implementing Project SEARCH in the context of an evaluation given that it has already been implemented in various locations and settings across the country. The program has a well-developed model, and each implementing site receives critical support at the beginning from the

national headquarters and continual guidance as needed. Regarding expectations about participant outcomes, as noted earlier, the program is focused on paid competitive employment for the youth, and has demonstrated that many of its participants actually achieve that goal. Further, a small experimental study focusing on a subgroup of participants—those with autism—found large impacts (Wehman et al. 2014). For each leading design we consider, we think that it would be possible to identify a credible and equivalent comparison group as required for an impact evaluation.

We discuss the details related to data availability and statistical power later when we describe the leading evaluation designs. Briefly, we anticipate that an impact evaluation of Project SEARCH would rely on data from administrative records from VR and the Social Security Administration (SSA). Regarding statistical power, we consider the ability to detect impacts precisely against the fact that each Project SEARCH cohort is relatively small, and conclude that with relatively small sample size, the magnitude of impacts one would be able to detect with confidence might be large but the program has the potential to achieve impacts of such magnitude.

B. Existing setting design: a quasi-experimental matched comparison group approach

The leading evaluation design under the existing setting, which we refer to as the "existing setting design," involves a quasi-experimental matched comparison group approach. The idea involves using existing data to identify comparison communities and youth that are similar to those served by Project SEARCH. In this approach, we would first create a pool of potential comparison areas (for example, counties, metropolitan statistical areas [MSAs], zip codes) where Project SEARCH is not offered but are in geographic proximity and have similar characteristics to those that are served by Project SEARCH. We would then use data from VR administrative records to identify high school students from both types of areas (that is, areas served and not served by the program) who meet the basic eligibility criteria to participate in Project SEARCH. Those in served areas would include Project SEARCH enrollees, but would also include other individuals because Project SEARCH does not enroll all youth who meet the program's basic eligibility criteria. We would define the served areas and the eligibility criteria as tightly as we can—to ensure that the percentage of those enrolled in Project SEARCH from the sample for the served areas is as large as feasible. The higher that percentage, the easier it will be to detect an impact of a given size.

More specifically, under this evaluation approach the intervention group would be comprised of students who reside in areas where there is a Project SEARCH site and would likely be eligible for the program. The intervention group students would need to be in the catchment area of a high school that is collaborating with Project SEARCH.[1] These students would be individuals with intellectual and developmental disabilities in their last year of high school and that have an IEP. These students often have some sort of prior employment

[1] We would attempt to determine the high school at which an individual is enrolled using VR data. At a minimum, we could use information in the national VR database (the RSA-911 file) on the youth's zip code and county to back out the likely associated public high school. Because high school catchment areas sometimes overlap, in some instances it might be difficult to determine which students are eligible to enroll in intervention high schools.

experience, can travel independently, and have strong support of their family. In this design, the evaluation would use existing administrative data from VR to explore the characteristics most common to those that participate in Project SEARCH. In addition, the students that actually participate in the program can be identified in the VR data, where there is either a direct indicator or a billing code for Project SEARCH participants.

The first step in identifying a comparison group under this evaluation design would involve choosing potential comparison areas (counties, MSAs, or zip codes) that are not served by Project SEARCH (that is, none of the high schools serving the area are collaborating with Project SEARCH) but are otherwise similar to the areas served by the program. We would select these potential comparison areas using information on geographic, socioeconomic, and service availability characteristics. Data on these area level characteristics can be gathered from the U.S. Census Bureau's ACS or the county database. Then, from these areas, we would define the set of students in VR data from which to draw the comparison group—that is, the ones who meet the basic criteria we identified for the intervention group students. Since school districts that offer Project SEARCH may display particular characteristics, we would also gather information on school districts to include in the analysis, such as the number of staff per pupil (by type of staff), educational programs offered, and the population of students. These data can be obtained from the U.S. Department of Education's Common Core of Data (CCD). We would then use propensity score matching—a statistical method to match on multiple factors (Rosenbaum and Rubin 1983)—at the individual level to select students in the comparison areas who are well matched on individual, community, and school district characteristics with program-eligible students in the intervention areas. For propensity score matching, we would focus on individual level characteristics such as age, grade, type of disability, family socioeconomic background, any work experience, as well as school district and area-level characteristics.

To identify the impacts of the program, we would compare the outcomes of students in the intervention and comparison group. Because the intervention group comprises students eligible to participate in Project SEARCH (not only those who actually participate), the impacts estimated using this approach apply to all students who had an opportunity to participate in the program. In other words, the estimated impacts are for all students whom Project SEARCH intends to treat (ITT impacts), irrespective of their actual participation in program services. Policymakers and practitioners are likely to be interested in impacts on those who are treated—those who actually received Project SEARCH services—in addition to being interested in impacts on those whom the programs intend to treat. A simple approach to estimate program impacts on those who actually participated in Project SEARCH (the so-called impacts of treatment on the treated, or TOT impacts) involves dividing the ITT impact estimates by the proportion of students who actually received Project SEARCH services (Bloom 1984). A more rigorous but complicated approach would involve matching comparison group students to actual Project SEARCH participants, which we discuss in more detail in Section D below.

There are some potential challenges in using the evaluation design. One potential challenge is that we would have to be able to sufficiently narrow down the set of students likely to participate in Project SEARCH using the available data so that the fraction that actually participate in the Project SEARCH sites is fairly high. This is important because if too small fraction of students in the intervention group actually participate, it will lower our ability to precisely estimate impacts of reasonable magnitude. We would try to address this concern by

clearly defining the target population of interest for Project SEARCH so that the fraction actually engaged with the program is not too small. Another challenge is that this method rests on the assumption that the propensity score matching would be able to account for all self-selection into being an eligible student in a community and school district where there is a Project SEARCH site.[2] If there are remaining unobserved factors that affect both being eligible for program services in an intervention area and the student outcomes, then the estimated impacts will be biased. To address this potential limitation, the propensity score model would have to be sufficiently rich to capture the selection process for Project SEARCH well. We would also have to avoid situations where the school(s) with which Project SEARCH collaborates are "exceptional" in important respects (for example, the quality of their special education programs) relative to potential comparison schools. Accounting for school district characteristics would help us mitigate this concern.

These challenges notwithstanding, the existing setting design offers some key benefits. First, it creates an opportunity to include a large number of Project SEARCH sites across multiple states, which will help improve the evaluation's ability to detect smaller impacts. Second, by covering a large number of program sites, the design would create the basis for drawing conclusions about Project SEARCH's impacts that are roughly generalizable to the broader population that the program serves across states. Third, by focusing on those who are eligible to participate in Project SEARCH (instead of actual participants), it allows a way to account for some unobserved selection into participation among the broader group of eligible students.

C. Demonstration setting design: a randomized experimental approach

The leading evaluation design in the demonstration setting, which we refer to as the "demonstration setting design," involves an RCT. In this approach, we would work with school districts or LEAs and get informed consent from students to enroll in an evaluation. Then we would randomly assign students who enrolled in the demonstration to either a treatment group in which they have the opportunity to apply to Project SEARCH, or to a control group in which they are offered the usual transition services from the state VR agency. Note that in this design, the Project SEARCH partners (the host business, the teacher, VR representative, etc.) still have the chance to decide who actually participates in Project SEARCH; the randomization only narrows the pool of potential applicants. With an RCT, the gold standard for impact evaluation, we expect the design to produce unbiased impacts of having the opportunity to participate in Project SEARCH.

Using this design, we would estimate ITT impacts but would also be able to estimate TOT impacts rigorously. Because this design would cover students whom Project SEARCH intended to treat (not just those who actually participate in the program), it would produce estimates of ITT impacts. To generate impacts on those who actually received Project SEARCH services (TOT impacts), we could use an instrumental variables approach with the randomly assigned treatment status as an instrument for actual participation (Angrist et al. 1996). This approach would utilize the exogenous increase in the likelihood of participating in Project SEARCH resulting from random assignment to the treatment group to estimate the average impact on those

[2] For example, if parents move (or adjust the student's address) so their children can be in a school served by Project SEARCH, it might be challenging to account for such self-selection. This may not be a significant concern as it wasn't mentioned by any of the Project SEARCH staff or other partners during our site visits.

who participated in the program. The Bloom (1984) adjustment of dividing the ITT impacts by the proportion of treatment group member who received Project SEARCH services would be a simpler (but less rigorous) way of deriving the TOT impacts.

The primary challenges of this evaluation design would be getting the support of local school districts to agree to randomize individual students and to obtain a sufficiently large number of students who agree to participate in the evaluation. In a situation where impacts might be very large, we could start the trial with a small sample size, then expand—by involving more schools over multiple years—as needed to estimate impacts with sufficient precision to satisfy the need for producing adequate information for practitioners and policymakers.[3] It is likely to take some time and resources to find appropriate partners, to obtain the IRB approvals, and to get informed consent for a large number of participating students. At the same time, this evaluation design offers the most feasible approach out of the set of alternative designs we have considered for obtaining rigorous impact estimates. It would provide valuable information not only on the effectiveness of the program, but would also inform our broader understanding of what may or may not work to integrate transition-age youth into the labor market and their communities more broadly.

D. Alternative evaluation designs under the existing setting

In this section we describe other evaluation designs we have considered for a rigorous impact evaluation of Project SEARCH under the existing setting. In addition to outlining the general approach and data sources for each alternative evaluation design, we describe its potential drawbacks or challenges to implementing the design.

1. Matched comparison group with VR youth not in Project SEARCH

An alternative quasi-experimental design we considered is to construct the comparison group from matched VR participants not in Project SEARCH who reside in the same or neighboring geographic areas as the Project SEARCH participants. Unlike the leading existing setting design, the intervention group under this alternative design would consist of VR eligible students who actually participate in Project SEARCH (instead of all VR eligible students who meet the basic Project SEARCH eligibility criteria). In addition, this alternative design does not entail a deliberate, thoughtful choice of the comparison areas and school districts where comparison students reside. Because VR eligibility is a criteria for Project SEARCH participation, looking for potential comparison students among the VR enrolled youth seems appealing. Implementation of this approach would require close collaboration with the state VR agencies, as the evaluation would need access to VR data to identify Project SEARCH participants and the potential comparison group students in the VR data. From discussions with VR staff in Florida, Ohio, and Kentucky, our understanding is that we would be able to identify Project SEARCH participants in the VR data either through an indicator in the VR records or

[3] Wehman et al. (2014) demonstrated the applicability of a randomized experimental evaluation of Project SEARCH with a relatively small sample size of 44 students. The evaluation design we are discussing here would be different from what Wehman et al. (2014) did in a few ways: we would not restrict it to students with a specific type of disabling condition (versus their focus on students with Autism Spectrum Disorders); control group students in the proposed design would be offered standard VR services (versus following their IEPs); and lastly, we propose to expand the evaluation to a larger scale, with the small scale evaluation only as a starting point.

through a special billing code used for Project SEARCH participants. Therefore, we would be able to use VR data on individual characteristics (such as age, gender, disabling condition, and household income) to conduct propensity score matching and find a comparison group of VR enrollees that matches with Project SEARCH participants. To estimate impacts of Project SEARCH, we would compare outcomes for students in Project SEARCH and those in the matched comparison group.

Even though this approach seems similar to the leading existing setting design, it would produce estimated impacts of actual participation in Project SEARCH (not the impact of the opportunity to participate in the program). Conceptually that seems appealing, but the challenge would be to account for an additional layer of selection that goes with actual participation in Project SEARCH. This approach assumes that we will be able to account for selection into program participation by matching on observable individual characteristics. However, there might be unobserved factors behind individual student's decision to participate (or not participate) in the program, and these factors may also influence the student's employment outcomes. Examples include family support, student ability and motivation, LEA characteristics, and economic opportunities for youth. Because we won't be able to account for such unobserved differences, comparing outcomes for participants and the matched comparison group would not produce unbiased estimates of impacts of participation in Project SEARCH.

2. Instrumental variables approach

We have also considered an instrumental variables approach for the evaluation. One possible instrument is the number of Project SEARCH applicants in a year relative to the number of openings, which is typically fixed and decided in advance based on the partners and the funding available. This approach would involve using VR data in conjunction with data from Project SEARCH on the number of spaces and applications in a given year. The primary concern with this approach is that the instrumental variable might not meet the *exclusion restriction*, because more applicants in a year relative to Project SEARCH slots might imply greater competition for jobs in later years among members of that cohort. We have also thought about other instrumental variables, such as the number of businesses within commuting distance for a VR participant (for instance, 5, 10 or 15 mile radius of the participant's residence) that have more than 600 employees (which is typically the size needed for a Project SEARCH host business). However, we are not confident this would meet the *exclusion restriction* either because areas with a higher concentration of large businesses are likely to have greater employment opportunities for all youth with disabilities, not just for Project SEARCH participants.

3. Fuzzy regression discontinuity design using the rubric score

For evaluating the impacts of Project SEARCH, we also considered a regression discontinuity (RD) design, which is a rigorous impact evaluation method. The Project SEARCH guidelines note that students with a rubric score of 50 to 70 are ideal for the program. This provides a setting where individuals with just above 50 point score, who are more likely to be accepted into the program, can be compared to those with a score just under 50, who are less likely to be accepted into Project SEARCH. The appeal of the RD approach is that it creates a credible comparison group in a non-experimental setting because those just below the cutoff are likely to be very similar to those just above, arguably differing only in being accepted in Project SEARCH.

However, it is unclear whether we would see a jump in the probability of a student participating in Project SEARCH at 50 points. After conversations with the Project SEARCH leadership team, it seems that the 50 point cutoff score is not always used in practice. If there are many spaces to fill and few applicants in a year, the recruitment team may move down below 50 on the rubric scale when accepting applicants. Similarly, in years when there are many applicants, they may accept individuals well above the 50 cutoff. A fuzzy RD design approach could potentially address that there are some individuals in the program group below 50 and some individuals in the comparison group above 50. However, the concern is whether we would indeed see a jump at 50 in the probability of being in Project SEARCH. If we do not, we cannot credibly compare outcomes of those above and below 50 to identify the impacts of Project SEARCH.

There are a few other concerns in adopting this approach in practice. First, we need a relatively large number of cases around the threshold of 50 to carry out the analysis. Since Project SEARCH cohort sizes are relatively small, having a sufficiently large sample size required for effectively implementing the RD approach could be problematic. Second, we would need data and identifying information for all Project SEARCH applicants if we were to use administrative data from VR or SSA on employment related outcomes of students in the study. This would involve going to each school district individually to seek approval to obtain the identifying information, which can be very burdensome and costly with no guarantee of obtaining such data.

4. Regression with indicators of rubric score categories

This approach involves conducting a regression analysis for all Project SEARCH applicants using the ordinary least squares approach with interactions of a Project SEARCH participation indicator and indicators of rubric score categories. This is similar to the RD design discussed above, but this approach would compare Project SEARCH applicants with similar rubric scores to one another. The advantage is that this approach is less demanding in terms of the sample size. An important drawback, however, is that interpreting the estimated relationships between Project SEARCH participation and student outcomes as causal impacts would require an assumption that differences in the likelihood of being accepted in Project SEARCH within a given rubric score category is not influenced by any other unobserved factors. We have reasons to believe that this assumption is unlikely to hold as Project SEARCH selection committee does consider other factors (such the student's and their family's motivation and outlook about competitive employment). In addition, similar to the RD design, we would require identifying information for all Project SEARCH applicants (including non-participants).

5. Special education students from local school districts who do not participate in Project SEARCH as comparison group

One possible design we considered is to form a comparison group using students with disabilities who do not participate in Project SEARCH but are in the same school district as those served by the program. The advantage of this approach is that the comparison group is exposed to the same educational and employment environment and faces the same local labor market as those who participate in Project SEARCH. A major concern with this approach is that there are possible spillovers of Project SEARCH within the school district; during the site visits we conducted for this evaluability assessment, we learned that nearby schools and local districts

cooperate together on resource use and putting together Project SEARCH opportunities. Consequently, it would be difficult to identify a suitable comparison group with plausibly similar observed and unobserved characteristics but without being directly or indirectly affected by Project SEARCH. In addition, another challenge with this approach is that we would need data with identifying information and individual characteristics for the students from school administrative records. It can be very difficult to get such information from school districts due to privacy concerns.

A variant of the same evaluation design could be to use students with disabilities from nearby school districts to form a comparison group. However, we would face similar challenges to above in terms of needing data with identifying information and individual characteristics for the students from school administrative records. In addition, there remains the possibility of resource sharing across local school districts for students with disabilities, which could contaminate the comparison group.

6. High school diploma policy change in Florida

The last potential evaluation design for estimating impacts of Project SEARCH under the existing setting involves exploiting a recent policy change in Florida related to the high school diploma options available to students with disabilities. Prior to 2014, students with disabilities had the option of receiving the standard high school diploma or a special diploma. The special diploma option focused on helping youth transition from school to the workplace or post-secondary education. Beginning in 2014, however, the special diploma was repealed and only the standard diploma was offered. Students with disabilities can now choose among three options: (1) the standard diploma option that is available to all students; (2) an option that has slight modifications in courses from the standard diploma; and (3) an option for students with significant disabilities that allows them to meet the requirements of the standard diploma with access courses and alternate assessment.[4]

Discussions with administrators in the transition offices of a Florida school division suggest that the change in diploma options may result in a different mix of students who participate in Project SEARCH. In particular, it is anticipated that fewer of the high functioning students with disabilities will continue to participate in the program because there is a greater emphasis on meeting academic requirements than in the past to obtain a diploma. This makes room for students who previously may not have been selected to participate in Project SEARCH to have an opportunity to do so.

The possible change in the composition of students who could apply to Project SEARCH is potentially captured by a unique administrative exercise conducted by the Orange County school district in Florida. The Orange County school district administers a rubric guide to all students with disabilities, and the rubric scores inform their decision to refer the students to specific

[4] Access courses are designed to provide students with significant cognitive disabilities access to general curriculum through accommodations, education, and consultation. These courses are designed to meet the core intent of the standards that apply to all students in the same grade, but at reduced levels of complexity. The Florida Standards Alternate Assessment (FSAA) measures the performance of the students in access courses. This alternative assessment is designed for students whose participation in the general statewide assessment program is not appropriate, even with accommodations.

transition support programs (including Project SEARCH).[5] Assuming that the rubric guide used by the school district didn't change over time, if the change in the diploma policy changes the mix of students who could apply to Project SEARCH, it would be captured by their rubric score: in the post-policy change period, these students are likely to have lower rubric score, on average, than those who were eligible to apply to Project SEARCH before the policy change.

We considered using the variation due to the diploma policy change in the type of students who apply to Project SEARCH to identify impacts of Project SEARCH. Under this design, we would identify students who are eligible to apply to Project SEARCH after the policy change, determine their rubric score assigned by the school district, and then identify students from the pre-policy change period who had similar rubric score. Because the school district would not have referred students with lower rubric scores to Project SEARCH in the pre-policy change period but their counterparts in the post-policy change period are being referred to Project SEARCH, the former group could serve as a comparison group. One limitation of this approach is that there could be a cohort effect due to different mix of students in the pre-policy change cohort as well as due to other changes over time in the local economic and educational policy environment. To address that, we could use a third group of students with disabilities with very high or very low rubric scores (such that they would not be referred to Project SEARCH) from the post-policy change period as a second comparison group. Contrasting outcomes across the three groups (in a difference-in-difference-in-differences framework) may allow us a way to estimate impact of the opportunity to participate in Project SEARCH.

A disadvantage of this approach is that it is narrower in scope in that it focuses on estimating impacts for a particular set of individuals who are not currently being targeted by the program, many who would have had access to other transitional programs. Another potential challenge is that because of the change in the diploma policy for students with disabilities, transition teachers may assess students differently and assign different values for items in the rubric guide. Thus, even if the rubric guide remains the same, the scores may be confounded by the diploma policy change, in which case the evaluation approach would not work. In addition, due to privacy concerns, it may be difficult to get the necessary identifying information and individual characteristics for students from the school district.

E. Alternative evaluation designs under the demonstration setting

In this section we describe two alternative evaluation designs we have considered for a rigorous impact evaluation of Project SEARCH under the demonstration setting. In addition to outlining the general approach and data sources for each design, we describe its potential drawbacks or challenges to implementing the design.

1. RCT with youth eligible for VR services

In a demonstration setting, one evaluation design we considered is an RCT with youth who are eligible to receive VR services. Under this design, high school aged youth who apply for and are determined to be eligible to receive VR services are randomly assigned either to a treatment

[5] Our impression from discussions with administrators in transition offices in Florida is that the youth's residential location is not a key consideration for referring students to specific transition support programs (Project SEARCH or others).

group that gets an opportunity to apply and participate in Project SEARCH, or to a control group that receives the usual VR services. Youth assigned to the treatment group would be more likely to apply to Project SEARCH than those in the control group. This approach is similar to the leading demonstration setting design: the Project SEARCH selection committee would still have the chance to decide who actually participates in the program, and the approach would produce ITT impacts. The key difference between this and the leading demonstration setting design is that random assignment would be performed with VR eligible youth (not with students in targeted school districts). With an RCT, we expect the design to produce unbiased impacts of having the opportunity to participate in Project SEARCH versus usual VR services.

This approach would be particularly useful in states where the VR agency is focused on expanding services to youth with disabilities. The Workforce Innovation and Opportunity Act (WIOA) of 2014 expands the vision for supporting transition-age youth with disabilities, and it would require 15 percent of public VR agency funds to be used for pre-employment transition services to high school students. From our discussions with VR agency staff during the site visits for the current evaluability assessment, expanding services to youth through Project SEARCH to meet the WIOA requirements is an idea in which VR agencies would be interested. However, we understand that it might be quite challenging for the VR agencies to participate in an RCT due to reluctance in withholding information about Project SEARCH to a group of individuals. Further, school administrators suggested that most youth with disabilities are aware of Project SEARCH in those areas where it is available. As such, it is uncertain the extent to which the referral through VR would make a difference in participation in the program. To address this issue, Project SEARCH would have to agree to assess applications with VR referrals only, at least for the cohort of students involved in the evaluation. It's not clear how difficult or easy it would be to have the VR agencies and Project SEARCH agree to these requirements. In particular, the VR agencies may have ethical concerns as they might consider that by not referring the comparison group to Project SEARCH they would be taking something away from them—an opportunity that is standard and available to individuals in the area with similar needs and employment goals.

2. Phased expansion of Project SEARCH sites with WIOA

As discussed above, VR agencies are considering the possibility of using Project SEARCH as a means to meet the WIOA requirement for extensive pre-employment transition services for youth with disabilities. In particular, there has been discussion about expanding the number of Project SEARCH sites to provide services to greater numbers of transition-age students. If this occurs, this could create an opportunity for a rigorous impact evaluation where Project SEARCH sites are rolled out in phases. For example, if VR is considering expanding two Project SEARCH sites, one possibility would be to randomly choose one site to start first, followed by the other in a couple of years. Then, students at the delayed site who do not yet have access to Project SEARCH (but would have had access if their site had been chosen first), could act as a comparison group for those at the first site. Since the order of expansion of the sites is random, then it mitigates concerns of selection into Project SEARCH at the school and community level.

A major challenge with this approach is that it requires the agreement and cooperation of many players (LEAs, VR, host businesses, etc.), specifically around the random phasing of the sites. Some players may not be open to delaying. Further, it is uncertain how generalizable this design would be: it analyzes a particular context where a new Project SEARCH site is introduced

to a community, where students previously didn't have the option to attend Project SEARCH, and in a time period where there may be growing pains given it is a new site. Given its narrower scope and the uncertain likelihood of implementing this approach, we feel it would be a challenging evaluation design to implement.

This page has been left blank for double-sided copying.

III. KEY COMPONENTS OF PROJECT SEARCH IMPACT EVALUATION

A. Objectives and research questions

The basic objective of a rigorous impact evaluation of Project SEARCH would be to produce evidence on whether the program is successful in facilitating (1) youth's receipt of employment promoting transition services that help youth successfully transition from secondary school to employment, and (2) improvements in youth's wellbeing through increased earnings and income. To attain these objectives, an impact evaluation of Project SEARCH would answer the following four research questions:

1. **Does Project SEARCH lead to receipt of more and better transition services for youth with disabilities?** The impact analysis would determine whether Project SEARCH increases the receipt of employment promoting transition services by youth relative to what they would have received in the absence of the program. This would include examining youth's engagement in paid or unpaid work during the year when youth are engaged with Project SEARCH.
2. **Does Project SEARCH improve employment related outcomes for youth with disabilities?** The impact evaluation would assess the success of the program in improving youth's employment related outcomes, such as hours of work, earnings, and income.
3. **Does Project SEARCH reduce youth's receipt of benefits from SSA disability programs?** If appropriate SSA administrative data are available for an evaluation, the impact analysis would be able to determine whether Project SEARCH reduces the youth's receipt of SSA disability program benefits. Even though Project SEARCH doesn't particularly focus on serving students who are SSA disability program beneficiaries, understanding the program's impact on receipt and amount of benefits for SSA beneficiaries would be of substantial policy interest.
4. **Is Project SEARCH more effective for some youth than for others?** With sufficient sample sizes, the evaluation could estimate program impacts on key subgroups of youth. These subgroups would be defined by characteristics before youth could engage with Project SEARCH, such as type of disabling condition, prior work experience, and receipt of disability program benefits at SSA. Comparisons of program impacts across these subgroups may help the program to tailor their efforts in the future to particular subsets of youth.

B. Outcome domains and key measures

In addressing the research questions, the evaluation would focus on short-term outcomes (such as service receipt) as well as medium-term outcomes (such as employment in paid jobs, hours of work, earnings, and income). In Table 1, we present a list of outcome domains for the impact analysis, and recommended measures within those domains. The outcome measures reflect types of outcomes that would be of interest for a medium term impact evaluation. Also, some of the outcomes can be constructed using data from administrative records at VR or SSA, but others may require a follow-up survey administered with youth involved in the evaluation.

The outcome domains shown in Table 1 are: service receipt, employment related outcomes, and youth well-being. Outcomes in the service receipt domain would allow the evaluation to

assess whether Project SEARCH makes a difference in terms of youth's receipt of transition services or youth gaining some work experience during the year the program group members engage with Project SEARCH. The latter outcome is a measure of service because Project SEARCH offers unpaid internships to all participants. If the program does not have any impact on outcomes in the service receipt domain, it would be hard to expect impacts on outcomes in the other domains.

Outcomes in the employment and youth well-being domains reflect Project SEARCH's goal of achieving competitive employment for all participants. Thus, the evaluation should assess whether the program made a difference in terms of youth having a paid job and working at least 20 hours per week, and to higher youth earnings. Youth income is a measure that would combine youth earnings with any cash benefits they may receive from public programs. Given the population served by Project SEARCH, receipt of assistance from the SSA disability programs would not be uncommon. With improved earnings, it is conceivable that the youth would rely less on public assistance—so we envision that the evaluation would assess impacts on related outcomes. The outcomes in the employment and youth well-being domains can be measured any time after the program group youth have completed their engagement with Project SEARCH. We envision that these outcomes would be measured at least a year after Project SEARCH services are completed, but depending on data availability they could be measured in the longer term—two, three, or five years after program completion.

Table 1. Suggested outcome domains and measures for a Project SEARCH impact evaluation

Domain and type of outcome	Measure
Service receipt	
Primary outcome	Receipt of employment related transition services
Secondary outcomes	Receipt of non-employment transition services (education, financial literacy, benefit counseling, other); early work-based experience during the year program group youth engage in Project SEARCH
Employment related outcomes	
Primary outcome	Employed in a paid job
Secondary outcomes	Hours of work; earnings
Youth well-being	
Primary outcome	Youth income
Secondary outcomes	Receipt of SSA disability program benefits; amount of SSA disability program benefits, participation in other public assistance programs

C. Data sources

For the evaluation designs discussed in this report, particularly for the leading designs under the existing and demonstration settings, we envision relying heavily on data from administrative records and other existing sources. In particular, we think that administrative records from the state VR agencies and from SSA could facilitate a rigorous evaluation of Project SEARCH without undertaking a large survey data collection effort. Relying on administrative records for youth outcomes would not only be cost efficient, but it would also create an opportunity to

examine impacts on youth employment and other outcomes several years after completion of the program—a task that is both problematic (due to attrition) and expensive when relying on survey data.

The existing setting design could use: the ACS for data on county characteristics (for example, county demographic mix in terms of age, race, gender, educational attainment; median income of the county, percentage of population with a disability; urban-rural status) or the Local Area Unemployment Statistics database from the Bureau of Labor Statistics to identify comparison areas. We would also consider using the CCD from the U.S. Department of Education's National Center for Education Statistics for data on schools (for example, the staff to pupil ratio by type of staff; the number of students; type of school). We would then rely on state VR administrative records to identify potential comparison group students by using individual-level data on location of residence, type of disabling condition, whether they have an IEP, as well as other characteristics to be used in the propensity score matching analysis. For data on youth outcomes, we envision using data from VR records as well as from SSA administrative data. If we have individual identifying information, such as names, Social Security Number (SSN), and date of birth, it would be possible to identify the study sample in SSA administrative data files (which may require specific data use agreements with the agency).

The demonstration setting design would require buy-in from LEAs to participate in a randomized experimental evaluation and also affirmative consent from individual students to enroll in the evaluation. Once these agreements and consents are in place, we envision relying on VR and SSA data on student level characteristics, as well as student outcomes. As part of gathering individual consent to enroll in the evaluation, we could collect individual identifying information (that is, names, SSN, and date of birth) to facilitate identifying the study sample in the administrative data files. Once they are identified in the administrative records we would be able to use pre- and post-program information from the administrative data files for the impact analysis.

Finally, if necessary, the evaluation could undertake a follow-up survey. Even though conducting a survey would require substantial resources, it would create an opportunity to collect information on a large variety of youth characteristics and outcomes beyond what would be available from administrative records. These include data on the youth's educational outcomes (for example, school enrollment status, educational attainment), health status and health insurance coverage, contact with the justice system, living arrangements, and receipt of public assistance. The costs and benefits of collecting survey data will have to be weighed appropriately. The timing of the follow-up survey relative to program completion will have to be determined as well; the survey can be conducted one year after program completion, or at a later point if there is greater interest in longer term impacts of Project SEARCH. As time passes, it will become harder to find the subjects, so attrition will be higher.

This page has been left blank for double-sided copying.

IV. IMPACT ESTIMATION APPROACH UNDER THE LEADING DESIGNS

A. The existing setting design: a matched comparison group approach

1. Impact analysis

Under the existing setting design, the evaluation would identify the analytical sample using propensity score matching at the individual level, taking into account community, school district, and individual characteristics of students in the study. To derive impact estimates, we would compare the outcomes for students in the program group and their matched comparisons. The impact estimates would be derived by estimating regression models of the following form:

$$(1) \qquad y_{i,r} = \propto + \beta I_i + X_i + M_{i,r} + \epsilon_{i,r}$$

where for individual *i* who resides in area *r*, $y_{i,r}$ is the outcome of interest (for example, received transition services, employed in a paid job); I_i is an indicator for whether the individual resides in a community where Project SEARCH is offered; X_i are student characteristics such as gender, type of disabling condition, age, and other pre-program characteristics; $M_{i,r}$ are characteristics of areas and school districts included in the study (for example, local unemployment rate); and $\epsilon_{i,r}$ is a random disturbance term.

The key parameter of interest in this evaluation design is β, which measures the impact of being in an area served by Project SEARCH on key outcomes, taking into account individual and community characteristics. Since the evaluation design uses an ITT framework and the study sample would include students who reside in the areas served by Project SEARCH but who do not actually participate in SEARCH, we would divide the estimated β by the participation rate in our sample to derive the impact of participating in Project SEARCH.

Because the detailed specification under this approach involves many somewhat arbitrary decisions, such as the characteristics of students included in the evaluation before and after matching the evaluation would investigate the robustness of the impact estimates to those decisions. After performing matching using the propensity score matching method, the characteristics of students should look similar between those who reside in the program and comparison areas. We would also verify that there is common support and if the distribution of propensity scores are similar between the program and the comparison group.[6] Finally, we would check the sensitivity of estimated impacts to the inclusion of some covariates in the propensity score model as well as various matching techniques.

2. Statistical power and precision

Even with a rigorous evaluation design, sample sizes must be large enough to provide sufficient statistical power to estimate impacts precisely in cases where the program produces impacts that are small but meaningful to policymakers or practitioners. As mentioned earlier,

[6] The common support condition ensures that individuals with the same observable characteristics have a positive probability of being both in the treatment and the comparison group. It essentially rules out the perfect predictability of being in the treatment group given the set of explanatory variables used in the propensity score model.

Project SEARCH sites typically enroll about 10 to 12 students each year. Consequently, it might be challenging to gather a large enough study sample that would enable the evaluation to detect small impacts. Pooling across multiple sites will increase the total sample size for the impact analysis and would help with statistical power.

In Table 2 we present the minimum impacts we would be able to detect in the existing setting design, relying on administrative or survey data on three key youth outcomes: (1) employment in paid jobs, (2) annual earnings, and (3) receipt of SSA disability program benefit. We calculate the minimum detectable impacts (MDI) for different overall sample sizes, and assuming two matched comparison group students for every student in the program group.[7]

Table 2. MDIs with the existing setting design

	Outcome		
Sample size	Employed in paid jobs	Annual earnings	SSA disability program benefit receipt
Assumed mean value of outcome for comparison group members	23%	$900	50%
1,500	6.6%	$474	7.9%
1,200	7.4%	$530	8.8%
900	8.6%	$612	10.2%
600	10.5%	$749	12.5%
300	14.9%	$1,060	17.7%
150	21.0%	$1,499	25.0%
75	29.7%	$2,120	35.3%

Notes: MDI calculations assume (1) a ratio of 1:2 for the number of intervention to comparison members, (2) a 95 percent confidence level with an 80 percent level of power, (3) a two-tailed test, (4) a reduction in variance of 10 percent owing to the use of regression models, (5) the 15 percentage of the variance of the treatment indicator (t) explained by the individual-level covariates, (6) standard deviations of annual earnings of $3,000, and (7) administrative data obtained on 100 percent of the sample.

The MDIs are sizeable relative to the mean values, become larger with smaller sample sizes, and show some variation across outcomes. For example, we would be able to detect program impacts of 6.6 percentage points or larger on paid employment with 500 intervention and 1,000 comparison students. This is about 29 percent of the mean employment rate (23 percent) for the comparison group. Put differently, if Project SEARCH's true impact is less than 6.6 percentage points, the estimated impact is unlikely to be statistically significant when the total sample size is 1,500. With an overall sample size as small as 75, we would be able to detect impacts on employment of about 30 percentage points or larger. Similarly, with an overall sample size of 1,500, the MDI for annual earnings is $474 (about 53 percent of the mean annual earnings of $900 for the comparison group), and with a sample size of 75, the MDI is $2,120 (about two and a half times the comparison group mean). For receipt of SSA disability program benefit, with an

[7] The MDI calculations reflect the fact that we envision matching at the individual-level, as noted in the main text. Even though we would first identify comparison areas, we would not formally match comparison and intervention areas, and instead use area-level characteristics in the individual-level propensity score matching analysis.

overall sample size of 1,500 and 75, the MDI is 7.9 and 35.3 percentage points, respectively. Considering the outcomes expected for Project SEARCH participants, the magnitude of the MDIs for paid employment and annual earnings are not large in absolute terms and may indeed be feasible. However, it might be more challenging for the program to reduce the student's receipt of SSA disability program benefits in the short- to medium-term. Thus, even though it might be feasible to detect statistically significant impact estimates on employment related outcomes with overall sample sizes as small as 300, the impact evaluation would have to incorporate larger sample sizes to be able to detect impacts of smaller magnitude on SSA disability benefit receipt.

As mentioned earlier, narrowing the definition of the eligible population so that the Project SEARCH participation rate in the intervention group is high will be critical for the ability to detect impacts of Project SEARCH. The MDIs shown in Table 2 are lower bound of the MDIs, as the calculations assume a 100 percent participation rate (that is, all students in the intervention group would participate in Project SEARCH). The ability of the impact analysis to detect impacts will decrease as fewer of the study sample actually participate in Project SEARCH services. For instance, if 40 of 100 students in the intervention group were to actually participate in the program instead of say 80 of 100, the minimum impact we could detect would double. Consequently, the lower the take-up rate of Project SEARCH among intervention group students, the worse would be the evaluation's ability to detect impacts.

B. The demonstration setting design: A randomized experiment

1. Impact analysis

The demonstration setting design involves randomly assigning students who agree to enroll in the evaluation into a treatment group, where they can apply to take part in Project SEARCH, or to a control group, where they cannot apply and instead have access to usual VR services. Given the RCT approach, we could estimate impacts of the opportunity to participate in Project SEARCH by computing simple differences in the mean values of outcomes between the treatment and control groups. However, more precise impact estimates can be obtained by estimating regression models of the following form:

$$(2) \qquad y_i = \propto + \beta T_i + X_i + M_i + \epsilon_i$$

where y_i is the outcome of individual *i* (for example, received transition services, employed in a paid job); T_i is an indicator for whether the individual is randomized into the group that can apply for Project SEARCH; X_i are student characteristics such as gender, primary impairment, age, and other pre-treatment characteristics; M_i are LEA and community-level characteristics (for example, percent of student receiving free or reduced price lunch); and $\epsilon_{i,r}$ is a random disturbance term.

Again, the key parameter of interest is β, which measures the impact of having the opportunity to apply for Project SEARCH. Since policymakers and practitioners are likely to be interested in impacts on those who actually participated in Project SEARCH services (that is, TOT impacts), the evaluation would be able to use an instrumental variables approach with the randomly assigned treatment status as an instrument for actual participation. Alternatively, the TOT impacts could be derived using the less rigorous but simpler Bloom (1984) adjustment—by

dividing the ITT impacts by the proportion of treatment group member who received Project SEARCH services.

As part of the impact analysis, we would assess whether random assignment resulted in two equivalent groups. To that end, we would examine whether the treatment and control groups are statistically equivalent on pre-program characteristics. We would also compare the results with and without regression adjustment to assess the sensitivity of the results; regression adjustment is most appropriate when the treatment and control group sizes are similar, as is the case we propose.

2. Statistical power and precision

As in the existing setting, because Project SEARCH sites typically enroll about 10 to 12 students each year, it might challenging to gather a large enough study sample that would enable the evaluation to detect small impacts. Table 3 shows the MDIs using the demonstration setting design. Even though the MDIs shown in the table are large relative to the control group mean, it is possible that Project SEARCH would result in impacts of such magnitude. The absolute magnitude of the impacts on the employment related outcomes are not always very large and existing evidence on Project SEARCH suggests that the program expects to produce large impacts. For example, with 1,000 students (500 treatment and 500 control), the MDI is just over $500 for annual earnings and 7 percentage points for being employed in a paid job. Given the evidence to date on Project SEARCH, impacts of at least this magnitude seem possible for these outcomes. However, as noted for the existing setting design, the program's impact on SSA disability program benefits might be relatively small in the medium-term. Consequently, having larger sample sizes would be necessary to enable the impact analysis to detect smaller impacts on benefit-related outcomes.

Table 3. MDIs with the demonstration setting design

	Outcome		
Sample size	Employed in paid jobs	Annual earnings	SSA disability program benefit receipt
Assumed mean value of outcome for control group members	23%	$900	50%
1,000	7.1%	$505	8.4%
800	7.9%	$564	9.4%
600	9.1%	$652	10.9%
400	11.2%	$798	13.3%
200	15.8%	$1,129	18.8%
100	22.4%	$1,596	26.6%
50	31.7%	$2,257	37.6%

Notes: MDI calculations assume (1) an equal number of treatment and control members, (2) a 95 percent confidence level with an 80 percent level of power, (3) a two-tailed test, (4) a reduction in variance of 10 percent owing to the use of regression models, (5) standard deviations of annual earnings of $3,000, and (6) administrative data obtained on 100 percent of the sample.

C. Other analytic issues

1. Multiple comparisons issue

Given that we propose to estimate impacts on several outcomes, we must be mindful of the statistical problem of "multiple comparisons." When researchers estimate impacts on a large number of outcomes, they are likely to see that at least a few of the estimates are probably statistically significant by chance, even if no true impacts occurred. We suggest taking a balanced approach to addressing the multiple comparisons problem (Schochet 2009) by making a tradeoff between reducing the likelihood of getting "false positives" (that is, finding statistically significant impacts by chance even when no true impacts exist) and maintaining our ability to avoid "false negatives" (that is, the statistical power to avoid incorrectly inferring no impacts when true impacts exist). First, we suggest determining a parsimonious set of outcome domains and specifying one (or a few) primary outcomes in each domain. We demonstrate this approach in Table 1 above. The primary outcomes would provide the basis for tests of the main hypotheses. By limiting the number of main hypotheses to be tested, this approach reduces the likelihood of finding impacts by chance alone, without significantly undermining the evaluation's statistical power to detect true impacts. Second, we suggest estimating impacts on additional outcomes in each domain (termed as secondary outcomes), but we would interpret the estimated impacts on the secondary outcomes cautiously, highlighting the findings only for the secondary outcomes if we found a credible pattern of statistically significant impacts on them.

2. Subgroup analysis

We suggest identifying subgroups of interest before conducting the impact analysis, and focusing on key characteristics of the youth that are policy relevant. In addition, the sample sizes of each subgroup would be a key factor in determining which groups of youth we analyze since we would require sufficient statistical power. Examples of subgroups we would explore include the type of disabling condition, prior work experience, and receipt of disability benefits. We would limit the number of subgroups and the outcomes to be evaluated, in order to minimize the risk of drawing spurious conclusions due to multiple comparisons.

This page has been left blank for double-sided copying.

V. CONCLUSION

In this report we presented several design options for an impact evaluation of Project SEARCH, a promising employment-related transition service program for high school students with disabilities. Relying on information we gathered from document reviews and from site visits conducted for this evaluability assessment, we have proposed two leading evaluation designs. First, in the existing setting design, we proposed a matched comparison group design with eligible youth from areas not served by Project SEARCH to be matched with those from areas served by the program. Second, in the demonstration setting design, we proposed a randomized experimental evaluation with school districts/LEAs randomly assigning youth enrolled in the demonstration either to a treatment group that would have the opportunity to apply for Project SEARCH services, or to a control group that would have the opportunity to receive usual services from the VR agency. We also briefly discussed other alternative design options, as well as details of implementing the leading design options.

We recommend the pursuit of the existing setting design as the preferred option, for practical reasons. We believe this design would meet the standards or rigor necessary for the findings to be credible to policymakers and other stakeholders, and importantly, would be the most feasible to implement. In particular, it would use existing Project SEARCH sites and participants for the intervention group and would also use existing data sources (state VR, SSA, ACS, CCD). Not only would these design features contain the costs of an evaluation, but they would also minimize some of the uncertainties and upfront coordination necessary to implement a design in the demonstration setting. Nevertheless, this design would still require careful planning and substantial effort. Specifically, researchers must determine exactly which Project SEARCH sites to focus on and the years of study; obtain research agreements with participating state VR agencies to access and link their data to SSA data; and gather data from other data sources to identify a credible comparison group.

Even though the demonstration setting design offers what many would consider to be the gold standard for impact evaluations, an RCT, it also would involve significantly more resources and a longer time frame to implement. In this design, school districts would randomly assign youth either to a treatment group that would have the opportunity to apply for Project SEARCH services or to a control group that would have the opportunity to receive usual VR services. As such, researchers would need strong collaboration and buy-in from the local school districts and would have to obtain informed consent from each participant. Further, this design would face more stringent IRB approval requirements, and researchers would be required to collect and store at least some data from participants. Sample sizes would likely be smaller than under the existing setting design, making it more difficult to detect smaller impacts. These steps all involve significant resources and time to implement, leading to some uncertainty about its feasibility and success.

As in all social program evaluations, there are various threats under both the existing and demonstration setting design that could undermine the evaluation's ability to draw meaningful conclusions for policymakers and other stakeholders. In the existing setting design, the key threat is related to "internal validity": there may be unobservable factors that drive the differences in outcomes between the intervention and the matched comparison group and consequently bias the impact estimates. To minimize this threat, the researcher must use a rich set of community,

school district, and individual characteristics in the analyses. The demonstration setting design also poses some threats despite relying on an RCT. In particular, this design is likely to have relatively small samples sizes and be conducted in a small number of settings, making it harder to generalize evaluation findings to other settings—an external validity problem. An additional potential threat is that youth randomly assigned to the control group may end up receiving Project SEARCH services; such contamination would threaten internal validity. Furthermore, since students from the same school districts would be randomly assigned to treatment and control groups, the treatment group members may not get the cohesive institutional supports they would receive in the standard Project SEARCH setting, which might also reduce the external validity of the evaluation findings. These potential threats to both the existing setting design and the demonstration setting design may limit the usefulness of results from an impact evaluation of Project SEARCH to policymakers and other stakeholders. However, the possibility of these threats materializing is not unusually or exceptionally high, and not apparently larger for one design than the other. Given that there is some risk in pursuing either of these designs helps tip the scale in favor of the design that uses far less resources, the existing setting design.

Implementing either of the leading evaluation designs would require collaboration with Project SEARCH and other entities. If an impact evaluation of Project SEARCH is pursued, we envision that it would require further discussion with the Project SEARCH leadership team as well as site staff. In addition, researchers involved in the evaluation would need buy-in from partners in each Project SEARCH site. Establishing a research and data use agreement with the state VR agencies and with SSA would also be required. Arranging additional funding would be necessary if a follow-up survey is administered. Once all data are available, the evaluation should produce a comprehensive report describing the program and comparison group conditions, as well as presenting the estimated impacts of the program.

REFERENCES

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association*, vol. 91, no. 434, June 1996, pp. 444–455.

Balcazar, F., T. Taylor-Ritzler, S. Dimpfl, N. Portillo-Pena, A. Guzman, R. Schiff, and M. Murvay. "Improving the Transition Outcomes of Low-Income Minority Youth with Disabilities." *Exceptionality*, vol. 20, 2012, pp. 114–132.

Bloom, H. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review*, vol. 8, 1984, pp. 225–246.

Butterworth, J., and A. Migliore. "Trends in Employment Outcomes of Young Adults with Intellectual and Developmental Disabilities, 2006-2013." Boston, MA: University of Massachusetts Boston, Institute for Community Inclusion, 2015.

Butterworth, J., F.A. Smith, A.C. Hall, A. Migliore, J. Winsor, and D. Domin. "StateData: The National Report on Employment Services and Outcomes." Boston, MA: University of Massachusetts Boston, Institute for Community Inclusion, 2014.

Carter, E.W., D. Austin, and A.A. Trainor. "Predictors of Postschool Employment Outcomes for Young Adults with Severe Disabilities." *Journal of Disability Policy Studies*, vol. 23, 2012, pp. 50–63. doi: 10.1177/1044207311414680.

Carter, E., A. Trainor, N. Ditchman, and L. Owens. "A Pilot Study Connecting Youth with Emotional and Behavioral Difficulties to Summer Work Experiences." *Career Development for Exceptional Individuals*, vol. 34, 2011, pp. 95–106.

Carter, E., A. Trainor, A. Ditchman, B. Swedeen, and L. Owens. "Evaluation of a Multi-Component Intervention Package to Increase Summer Work Experiences for Transition-Age Youth with Severe Disabilities." *Research and Practice for Persons with Severe Disabilities*, vol. 34, 2009, pp. 1–12.

Christensen, J., S. Hetherington, M. Daston, and E. Riehle. "Longitudinal Outcomes of Project SEARCH in Upstate New York." *Journal of Vocational Rehabilitation*, vol. 42, 2015, pp. 247–255.

D'Alonzo, B.J. "Career Education for Handicapped Youth and Adults in the '70s." *Career Development for Exceptional Individuals*, vol. 1, 1978, pp. 4–12. doi: 10.1177/088572887800100102.

Daston, M., J.E. Riehle, and S. Rutkowski. *High School Transition that Works: Lessons Learned from Project SEARCH*. Baltimore, MD: Brooks Publishing, 2012.

Fraker, T., E. Carter, T. Honeycutt, J. Kauff, G. Livermore, A. and Mamun. "Promoting Readiness of Minors in SSI (PROMISE) Evaluation Design Report." Washington, DC: Mathematica Policy Research, June 2014a. Retrieved from http://www.socialsecurity.gov/disabilityresearch/documents/PROMISE_Eval%20%20Design%20Report_Final.pdf

Fraker, Thomas, Todd Honeycutt, Arif Mamun, Allison Thompkins, and Erin Jacobs Valentine. "Final Report on the Youth Transition Demonstration Evaluation." Washington, DC: Mathematica Policy Research, November 2014b.

Halpern, A.S. "Transition: A Look at the Foundations." *Exceptional Children*, vol. 51, 1985, pp. 479–486.

Hemmeter, Jeffrey. "Earnings and Disability Program Participation of Youth Transition Demonstration Participants After 24 Months." *Social Security Bulletin*, vol. 74, no. 1, 2014, pp. 1–25.

Madaus, J.W., N.W. Gelbar, L.L. Dukes, M. Faggella-Luby, A.R. Lalor, J.S. Kowitt. "Thirty-five Years of Transition Topics: A Review of CDTEI Issues from 1978 to 2002." *Career Development and Transition for Exceptional Individuals*, vol. 36, 2013, pp. 7–14. doi:10.1177/2165143413476734.

Müller, E., and R. VanGilder. "The Relationship Between Participation in Project SEARCH and Job Readiness and Employment for Young Adults with Disabilities." *Journal of Vocational Rehabilitation*, vol. 40, 2014, pp. 15–26.

National Alliance for Secondary Education and Transition. "National Standards and Quality Indicators: Transition Toolkit for Systems Improvement." Minneapolis, MN: University of Minnesota, National Center on Secondary Education and Transition, 2005.

National Collaborative on Workforce and Disability for Youth. "Guideposts for Success." 2nd ed. Washington, DC: Institute for Educational Leadership, 2009.

Rosenbaum, Paul R., and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, vol. 70, no. 1, 1983, pp. 41–55.

Schall, Carol. "Project SEARCH with ASD Supports: A Randomized Clinical Trial to Explore Competitive Employment for 18 to 22 Year-Olds with Autism Spectrum Disorders (ASD)." Research Brief. Richmond, VA: Virginia Commonwealth University, 2013.

Schochet, Peter Z. "An Approach for Addressing the Multiple Testing Problems in Social Policy Impact Evaluations." *Education Review*, vol. 33, no. 6, December 2009, pp. 539–567.

Test, D., V. Mazzotti, A. Mustian, C. Fowler, L. Kortering, and P. Kohler. "Evidence-Based Transition Predictors for Improving Post School Outcomes for Students with Disabilities." *Career Development for Exceptional Individuals*, vol. 32, 2009, pp. 180–181.

Wehman, P., C. Schall, J. McDonough, J. Kregel, V. Brooke, A. Molinelli, W. Ham, C. Graham, J. Erin Riehle, H. Collins, and W. Thiss. "Competitive Employment for Youth with Autism Spectrum Disorders: Early Results from a Randomized Clinical Trial." *Journal of Autism and Developmental Disorders*, vol. 44, no. 3, 2014, pp. 487–500.

Wehman, P., A.P. Sima, J. Ketchum, M.D. West, F. Chan, R. Leucking. "Predictors of Successful Transition from School to Employment for Youth with Disabilities." *Journal of Occupational Rehabilitation*, vol. 25, 2015, pp. 323–334. doi: 10.1007/s10926-014-9541-6.

This page has been left blank for double-sided copying.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research