Evaluation of
Employment Coaching for
TANF and Related Populations

Selecting and Testing Measures of Self-Regulation Skills Among Low-Income Populations

OPRE Report #2020-138 • November 2020

Selecting and Testing Measures of Self-Regulation Skills Among Low-Income Populations

OPRE Report #2020-138 • November 2020

Tim Kautz • Quinn Moore

Submitted to:
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services
330 C Street, SW
Washington, DC 20201
Project Officers: Hilary Bruck and Victoria Kabak

Contract/Task Number: HHSP233201500035I / HHSP23337018T

Mathematica Reference Number: 50327

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

Suggested citation: Tim Kautz and Quinn Moore (2020). *Selecting and Testing Measures of Self-Regulation Skills Among Low-Income Populations*, OPRE Report # 2020-138, Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Disclaimer: The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research, and Evaluation are available at www.acf.hhs.gov/opre.

Sign up for the OPRE Newsletter

Follow OPRE
on Twitter
@OPRE_ACF

Like OPRE's page
on Facebook
OPRE.ACF

Follow OPRE
on Instagram
@opre_acf

Connect on
LinkedIn
company/opreacf

OPRE

Mathematica
Progress Together

Contents

TABLES

FIGURE

Overview

INTRODUCTION

The ability to find, keep, and advance in a job depends on self-regulation skills in addition to education, work experience, and technical skills (Almlund et al. 2011). Self-regulation skills include the ability to finish tasks, stay organized, and intentionally control emotions and behaviors. Research has shown that these skills are essential to attaining goals and determining life outcomes, including those related to employment (Almlund et al. 2011). At the same time, facing poverty, and the many stresses that accompany it, makes it particularly difficult to exercise self-regulation skills in the moment (Mullainathan and Shafir 2014; Hamoudi et al. 2014). For adults, interventions such as coaching can both strengthen self-regulation skills and encourage their use (Kautz et al. 2014). Research has also shown that a variety of interventions can improve self-regulation skills among children and youth (Murray et al. 2016). The Self-Regulation and Toxic Stress Series, sponsored by the Office of Planning, Research, and Evaluation (OPRE) of the Administration for Children and Families in the U.S. Department of Health and Human Services, explores interventions that support self-regulation across the lifespan and across contexts, and communicates the potential of a self-regulation framework for strengthening prevention programs and human services.

In response to this research and framework, some employment programs, including some that are offered as part of the Temporary Assistance for Needy Families (TANF) program, use coaching and other strategies designed to strengthen and boost participants' use of self-regulation skills (Cavadel et al. 2016; Kautz et al. 2014). To assess the effectiveness of these strategies, evaluators need a way to measure self-regulation skills accurately. Accurate measurement is important not only for understanding the results of individual studies, but for the ability to synthesize findings across studies and across developmental age groups (Murray and Rosanbalm, 2017).

The selection and testing of measures of self-regulation skills for evaluating the impact of such programs present unique challenges and considerations. In this setting, measures of self-regulation skills need to: (1) be feasible to implement in evaluations of employment programs; (2) be appropriate given the participants' backgrounds (for example, survey items need to be understood by participants); (3) capture skills that could be influenced by the employment program and could affect employment outcomes; and (4) be reliable and valid in that they consistently capture the skills they were designed to measure. The appropriate criteria for testing the performance of measures also differ from those applied in clinical and other settings. For use in impact evaluations, measures need to be sufficiently accurate for researchers to detect differences in self-regulation skills between treatment and comparison groups.

To provide guidance on how to address these challenges and considerations, this report describes a practical approach for selecting and testing measures of self-regulation skills for use in impact evaluations. It complements three earlier works: (1) Cavadel et al. (2018), which encourages practitioners to consider measuring outcomes related to self-regulation skills; (2) Kautz and Moore (2018), which provides guidance to researchers on how to measure self-regulation skills in evaluation settings; and (3) the Self-Regulation and Toxic Stress Series (Hamoudi et al. 2014; Murray et al. 2014, 2015, 2016) that presents information on the relationship between stress and self-regulation, as well as evidence and implications for interventions designed to improve self-regulation among children and youth. We describe the experience of implementing this approach for the Evaluation of Employment Coaching for TANF and Related Populations, a national study of coaching programs aimed at improving employment outcomes for low-income populations sponsored by OPRE. We provide empirical findings from tests of the measures selected for this study and describe how the results guided our decisions about how we will use the measures in the impact evaluation. We also discuss how these decisions could differ for studies that use measures with different empirical findings.

PRIMARY RESEARCH QUESTIONS

This report addresses the following research questions:

1. Which criteria should evaluators use when selecting and testing measures of self-regulation skills in the context of impact evaluations of employment programs that serve low-income populations?

2. How do the measures we selected for this evaluation perform?

   a. To what extent do respondents complete survey questions that measure self-regulation skills?

   b. Do survey questions designed to measure the same self-regulation skill consistently measure that skill?

   c. Do different self-regulation measures capture different skills?

   d. Are the survey questions grouped with the appropriate self-regulation measures?

   e. What is the correlation, or relationship, between measures of self-regulation skills and employment outcomes and challenges?

PURPOSE

This report discusses issues related to selecting and testing measures of self-regulation skills in evaluations of employment programs for low-income populations. First, it presents an overview of criteria for selecting measures of self-regulation skills. Second, through a presentation of empirical evidence, this report demonstrates a process for developing and testing self-regulation measures in the context of an impact evaluation of employment coaching programs for low-income populations. Third, it discusses how the process could be adapted to other studies.

KEY FINDINGS AND HIGHLIGHTS

For three reasons, it is challenging to develop and test measures of self-regulation skills in evaluations of employment programs for low-income populations. First, such evaluations have feasibility constraints and need to minimize the burden on individuals. Second, most measures of self-regulation skills were developed for other purposes and might not be appropriate in these evaluations. Third, the standard criteria for assessing the performance of measures do not always apply to evaluations.

To meet these challenges, we recommend (1) basing self-regulation measures on existing measures as much as possible; (2) selecting a subset of items from existing measures to reduce burden; (3) using a set of both general measures of self-regulation as well as measures specific to the employment context; (4) pre-testing measures to ensure that they are appropriate given the background of respondents; and (5) assessing the reliability and validity of the measures but considering carefully whether standard criteria for acceptable levels apply. For this evaluation, we met these criteria by selecting and developing appropriate questions for a baseline survey that was administered when participants enrolled in the study and follow-up surveys that will collect data on participants' outcomes.

In this report, we provide empirical findings from tests of the measures selected for the baseline survey and describe how the results guided our decisions about how we will use the measures in the impact evaluation. These results demonstrate the feasibility of developing reliable and valid measures of self-regulation skills for use in impact evaluations of employment programs for low-income populations. We found:

- Both the baseline survey as a whole and each individual item met the criteria for adequate levels of response rates
- The self-regulation measures captured distinct skills (exhibited discriminant validity)
- The items designed to measure a given skill consistently measured that skill (were reliable)
- The grouping of items into different self-regulation measures fit the data well overall, suggesting that the items were grouped with the right measures (exhibited model validity)
- The self-regulation measures were correlated with contemporaneously measured employment outcomes as well as with perceptions of potential employment challenges (exhibited concurrent validity)

METHODS

The report includes:

- A brief review of self-regulation skills and criteria for selecting corresponding measures in evaluations of employment programs for low-income populations

- Tests of the selected measures, including:
  - A nonresponse analysis that revealed the extent to which respondents completed items
  - A confirmatory factor analysis that assessed the extent to which measures captured distinct skills and the grouping of items into skills were appropriate
  - An assessment of reliability that shed light on the extent to which a group of items designed to measure a given skill consistently measured that skill
  - A correlational analysis that examined the relationship between the measures of self-regulation skills and employment-related variables
- Discussion of methodological approaches that could apply to different types of studies

Executive Summary

The ability to find, keep, and advance in a job depends on self-regulation skills in addition to education, work experience, and technical skills (Almlund et al. 2011). Self-regulation skills include the ability to finish tasks, stay organized, and intentionally control emotions and behaviors. Research has shown that these skills are essential to attaining goals and in determining life outcomes, including those related to employment (Almlund et al. 2011). At the same time, facing poverty, and the many stresses that accompany it, makes it particularly difficult to exercise self-regulation skills in the moment (Mullainathan and Shafir 2014; Hamoudi et al. 2014). For adults, interventions such as coaching can both strengthen these skills and encourage their use (Kautz et al. 2014). Research has also shown that a variety of interventions can improve self-regulation skills among children and youth (Murray et al. 2016).

An applied contextual model for promoting self-regulation in children and youth targets the social ecological environment and provides a frame for thinking about how to support self-regulation across the lifespan and across contexts (Murray et al. 2019). This orientation is consistent with the work of the Self-Regulation and Toxic Stress Series, which was sponsored by the Office of Planning, Research, and Evaluation (OPRE) of the Administration for Children and Families in the U.S. Department of Health and Human Services to communicate the potential of a self-regulation framework for strengthening prevention programs and human services.

In response to this research and framework, some employment programs, including some that are offered as part of the Temporary Assistance for Needy Families (TANF) program, use coaching and other strategies designed to strengthen and boost participants' use of self-regulation skills (Cavadel et al. 2016; Kautz et al. 2014). To assess the effectiveness of these strategies, evaluators need a way to measure self-regulation skills accurately. The selection and testing of measures of self-regulation skills for evaluating the impact of such programs present unique challenges and considerations. Accurate measurement is important not only for understanding the results of individual studies, but for the ability to synthesize findings across studies and across developmental age groups (Murray and Rosanbalm 2017).

We describe a practical approach for selecting and testing measures of self-regulation skills for use in impact evaluations.

To provide guidance on how to address these challenges and considerations, we describe a practical approach for selecting and testing measures of self-regulation skills for use in impact evaluations. We describe the experience of implementing this approach for the Evaluation of Employment Coaching for TANF and Related Populations, a national study of coaching programs aimed at improving employment outcomes for low-income populations sponsored by OPRE. We focused on three programs that are part of this evaluation.

SELECTING MEASURES OF SELF-REGULATION SKILLS

For evaluating the impact of employment programs for low-income populations, measures of self-regulation skills need to (1) be feasible to implement in evaluations of employment programs; (2) be appropriate given the participants' backgrounds (for example, survey items need to be understood by participants); (3) capture skills that could be influenced by the employment program and could affect employment outcomes; and (4) be reliable and valid in that they consistently capture the skills they were designed to measure.

For the Evaluation of Employment Coaching, we met these criteria by selecting and developing appropriate questions for a baseline survey that was administered when participants enrolled in the study and a follow-up survey that will collect data on participants' outcomes. The measures covered four key self-regulation skills: (1) Self-Esteem; (2) Emotional Control & Self-Monitoring; (3) Goal-Setting; and (4) Task Monitoring, Planning, & Initiation. Each measure is based on three or more survey items designed to capture the associated self-regulation skill.

To help ensure that the measures of self-regulation skills were appropriate for use in the evaluation's impact analyses, we tested their performance in the baseline survey on two key dimensions.

TESTING MEASURES OF SELF-REGULATION SKILLS

To help ensure that the measures of self-regulation skills were appropriate for use in the evaluation's impact analyses, we tested their performance in the baseline survey on two key dimensions. First, we assessed the response rates and response patterns, which suggested whether the measures of self-regulation skills were representative of the study participants and whether the questions captured the full range of each self-regulation skill.

Second, we assessed reliability and validity, which provided evidence about whether the measures captured distinct skills related to employment outcomes and whether the evaluation can detect impacts on the skills:

- *Reliability* refers to the extent that an assessment tool produces consistent results. In this report, we focus on one form of reliability, internal consistency, which refers to the degree to which different items for a given measure produce similar results.
- *Validity* refers to the extent to which a measure captures what it is designed to measure. We assessed three types of validity. First, we assessed the extent to which separate measures capture different constructs (discriminant validity) by examining the correlations between measures of different self-regulation skills. If two measures are highly correlated, then they likely capture the same underlying construct. Second, we assessed whether the self-regulation measures are correlated with employment-related measures collected at approximately the same time (concurrent validity). If self-regulation skills are not related to employment outcomes, then changes in self-regulation skills would not be expected to influence employment outcomes. Third, we assessed the extent to which the grouping of items into skills fit the data well overall (model validity). If we found that the groupings fit the data poorly, it would suggest that some items might belong with a different self-regulation skill instead.

The results of our analyses demonstrate the feasibility of developing reliable and valid measures of self-regulation skills for use in impact evaluations of employment programs for low-income populations (Table ES.1).

Table ES.1. Summary of main analyses

Research question	Summary of results
To what extent do respondents complete items?	**Both the survey as a whole and each individual item met the criteria for adequate levels of response rates**. All (100 percent) study participants took the survey. Response rates for individual items were at least 97 percent.
Do items designed to measure the same skill relate to each other?	**The items designed to measure a given skill consistently measured that skill (were reliable)**. In the full sample, the estimated Cronbach's alpha (a measure of reliability) was 0.65 or above for all measures, meeting our target threshold. The estimated reliability differed somewhat across employment programs.
Do different self-regulation measures capture distinct skills?	**The self-regulation measures captured distinct skills (exhibited discriminant validity)**. For all samples, the correlation between pairs of skills was 0.72 or less, which met our criteria that distinct skills should demonstrate correlations below 0.80.
Do the self-regulation measures relate to other important variables?	**The self-regulation measures were correlated with contemporaneously measured employment outcomes as well as with individuals' perceived lack of employment challenges (exhibited concurrent validity**). Although these findings did not lend themselves to straightforward interpretation, the results suggested that the self-regulation measures may allow us to study whether the employment programs affect later outcomes through self-regulation skills.
Are the survey items grouped with the appropriate self-regulation measures?	**The grouping of items into different self-regulation measures fit the data well overall, suggesting that the items were grouped with the right measures**. Across all samples, estimated fit statistics met standard criteria for acceptable fit. The results were stable across employment programs, suggesting some level of generalizability.

I. Introduction

The ability to find, keep, and advance in a job depends on self-regulation skills in addition to education, work experience, and technical skills (Almlund et al. 2011). Self-regulation skills include the ability to finish tasks, stay organized, and intentionally control emotions and behaviors. Research has shown that these skills are essential to attaining goals and in determining life outcomes, including those related to employment (Almlund et al. 2011). At the same time, facing poverty, and the many stresses that accompany it, makes it particularly difficult to exercise self-regulation skills in the moment (Mullainathan and Shafir 2014; Hamoudi et al. 2014). For adults, interventions such as coaching can both strengthen these skills and encourage their use (Kautz et al. 2014). Research has also shown that a variety of interventions can improve self-regulation skills among children and youth (Murray et al. 2016).

An applied contextual model for promoting self-regulation in children and youth targets the social ecological environment and provides a frame for thinking about how to support self-regulation across the lifespan and across contexts (Murray et al. 2019). This orientation is consistent with the work of the Self-Regulation and Toxic Stress Series, which was sponsored by the Office of Planning, Research, and Evaluation (OPRE) of the Administration for Children and Families in the U.S. Department of Health and Human Services to communicate the potential of a self-regulation framework for strengthening prevention programs and human services.

In response to this research and framework, some employment programs, including some that are offered as part of the Temporary Assistance for Needy Families (TANF) program, use coaching and other strategies designed to strengthen and boost participants' use of self-regulation skills (Cavadel et al. 2016; Kautz et al. 2014).[1] To assess the effectiveness of these strategies, evaluators need a way to measure self-regulation skills accurately. This is important not only for understanding the results of individual studies, but for the ability to synthesize findings across studies and across developmental age groups (Murray and Rosanbalm, 2017).

The selection and testing of measures of self-regulation skills for evaluating the impact of such programs present unique challenges and considerations. In this setting, measures of self-regulation skills need to (1) be feasible to implement in evaluations of employment programs; (2) be appropriate given the participants' backgrounds (for example, participants understand the survey items and interpret them as intended); (3) capture skills that could be influenced by the employment program and could affect employment outcomes; and (4) be reliable and valid in that they consistently capture the skills they were designed to measure. The appropriate criteria for testing the performance of measures also differ from those applied in other settings, such as for clinical purposes. For use in impact evaluations, measures need to be sufficiently accurate for researchers to detect differences in self-regulation skills between treatment and comparison groups. In addition, the ability to detect such differences depends on other aspects of the evaluation, such as sample size.

[1] While not all designers of coaching interventions explicitly conceptualize the design in terms of self-regulation, by focusing on goal setting and collaborative interactions, coaching helps participants practice self-regulation skills.

To provide guidance on how to address these challenges and considerations, this report describes a practical approach for selecting and testing measures of self-regulation skills for use in impact evaluations.[2] We describe the experience of implementing this approach for the Evaluation of Employment Coaching for TANF and Related Populations, a national study of coaching programs aimed at improving employment outcomes for low-income populations sponsored by OPRE. We provide empirical findings from tests of the measures selected for this study and describe how the results guided our decisions about how we will use the measures in the impact evaluation. We also discuss how these decisions could differ for studies that use measures with different empirical findings.

The report addresses the following research questions:

1. Which criteria should evaluators use when selecting and testing measures of self-regulation skills in the context of impact evaluations of employment programs that serve low-income populations?
2. How do the measures we selected for this evaluation perform?
   a. To what extent do respondents complete survey questions that measure self-regulation skills?
   b. Do survey questions designed to measure the same self-regulation skill consistently measure that skill?
   c. Do different self-regulation measures capture different skills?
   d. Are the survey questions grouped with the appropriate self-regulation measures?
   e. What is the correlation, or relationship, between measures of self-regulation skills and employment outcomes and challenges?

We show that there are reliable and valid measures of self-regulation skills that may be used to evaluate employment coaching programs for low-income populations.

We show that there are reliable and valid measures of self-regulation skills that may be used to evaluate employment coaching programs for low-income populations. The measures selected for this evaluation met appropriate criteria for reliability and validity and were feasible to implement while minimally burdening study participants. The measures correlated with employment indicators, suggesting that they capture substantively meaningful constructs. These findings are relatively stable across various subpopulations and three programs that participated in the evaluation, suggesting that the measures might perform well in other contexts.

In Section II, we provide background on the measurement of self-regulation skills, outline the goals of the evaluation, and describe how we selected self-regulation measures that met those goals. In Section III, we discuss how we tested the measures of self-regulation skills, addressing response rates, reliability, and validity. We present our conclusions in Section IV.

[2] This report complements three earlier works: (1) Cavadel et al. (2018), which encourages practitioners to consider measuring outcomes related to self-regulation skills; (2) Kautz and Moore (2018), which provides guidance to researchers on how to measure self-regulation skills in evaluation settings; and (3) the Self-Regulation and Toxic Stress Series (Hamoudi et al. 2014; Murray et al. 2014, 2015, 2016) that presents information on the relationship between stress and self-regulation, as well as evidence and implications for interventions designed to improve self-regulation among children and youth

II. Selecting measures of self-regulation skills

We selected and developed measures of self-regulation skills through a process that balanced reliability and validity with the practical considerations of data collection in an impact evaluation of employment programs for low-income populations.

We selected and developed measures of self-regulation skills through a process that balanced reliability and validity with the practical considerations of data collection in an impact evaluation of employment programs for low-income populations.

A. OVERVIEW OF SELF-REGULATION SKILLS AND MEASUREMENT APPROACHES

In the absence of a universally accepted definition of "self-regulation," we followed Cavadel et al. (2016) by using the term to cover the broad set of skills that allow people to intentionally control their thoughts, emotions, and behaviors. We focus on three categories of self-regulation that are relevant to finding, keeping, and advancing in a job: (1) personality factors, which include motivation and self-esteem; (2) emotional skills, which include the ability to recognize emotions in others and modulate emotion appropriately; and (3) cognitive skills, which include the ability to plan, execute tasks, and set goals. These self-regulation skills complement each other and enable people to set, pursue, and attain goals, including those related to employment (Cavadel et al. 2016). For example, improved planning skills could help people to prioritize tasks at work and improved emotional skills could help people to communicate effectively with co-workers.

Measures of self-regulation skills differ in *how* they measure skills (their mode) and *what* they measure (their content). Measurement modes include (1) self-reports in a survey or interview that typically asks people about how they tend to behave; (2) an observer report in a survey or interview; (3) a performance task in which participants complete an in-person or computer-based activity that requires the use of a particular self-regulation skill[3]; and (4) administrative records about behaviors such as attendance. In the context of impact evaluations, there are tradeoffs between these modes because they differ in terms of cost, the burden on participants, and the skills that they capture.[4] As discussed later, given the goals and parameters of the Evaluation of Employment Coaching, we focused on self-reported measures. However, our process for developing and validating the measures applies to the other modes as well.

An important aspect of content is the extent to which measures capture skills in a specific setting. Measures of *contextualized behaviors* capture behaviors in a given setting. For example, the extent to which people complete tasks at work is a behavior in the context of work. Measures of *generalized behaviors* capture broader skills that apply across several settings (for example, the tendency to finish tasks in general). In the Evaluation of Employment Coaching, we included a mix of contextualized and generalized measures.

[3] For example, the "Stroop color and word test" is a task-based measure of inhibitory control (Stroop 1935). Respondents are presented with text that spells the name of one color but that is written in ink of a different color. They are then instructed to name the color of the ink rather than the color spelled out in the text. They might see the word "red" written in green ink, in which case the correct response is "green." The more accurately they name the color of the ink, the better is their inhibitory control.

[4] See Kautz and Moore (2018) for additional discussion on these points.

B. STUDY AND SAMPLE

The selection of appropriate self-regulation measures requires consideration of the context of data collection, the study's population, and the intended use of the measures.

1. Background on the study

In this report, we focus on the selection and testing of self-regulation measures used in an impact evaluation of several employment coaching programs that serve low-income populations (see Box 1 for an overview of the evaluation and the employment programs). These employment programs rely on trained coaches to help participants set individualized goals and to provide motivation, support, and feedback as the participants use self-regulation skills, which in turn will help them pursue those goals. In doing so, the programs aim to help the participants succeed in the labor market and move toward economic security. The evaluation used a randomized controlled trial (RCT). Participants in each program were randomly assigned to either a treatment group that was eligible to receive the program's coaching services or a control group that was not eligible for such services. Although the primary outcome measures are related to employment and self-sufficiency, the evaluation is also measuring self-regulation skills. Self-regulation skills are important outcomes because coaching is hypothesized to improve employment and self-sufficiency outcomes specifically by strengthening or increasing the use of these skills.

Box 1. Evaluation of Employment Coaching for TANF and Related Populations

To learn more about the potential of coaching to help TANF recipients and other low-income individuals reach economic security, the Office of Planning, Research, and Evaluation (OPRE) of the Administration for Children and Families is sponsoring an evaluation of employment coaching models. Using an experimental research design, the evaluation examines the effectiveness and implementation of coaching interventions that aim to help low-income individuals succeed in the labor market. The evaluation will examine the impact of coaching on self-regulation skills and the role of self-regulation skills in generating any impacts on employment outcomes.

The coaching models in the evaluation are:

- **Family Development and Self-Sufficiency program (FaDSS)** in Iowa. Under contract to the state, 17 local human services agencies use grants from the Iowa Department of Human Rights to provide TANF recipients with coaching during home visits. Seven of those 17 agencies are participating in the evaluation.
- **LIFT** in Chicago, Los Angeles, New York City, and Washington, D.C. LIFT is a nonprofit organization that provides career and financial coaching to parents and caregivers of young children. LIFT sites in Chicago, Los Angeles, and New York City are participating in the evaluation.
- **Goal4 It!™** in Jefferson County, Colorado. Goal4 It!™ is an employment coaching intervention designed by Mathematica and partners that is being piloted in a TANF program as an alternative to more traditional case management.
- **MyGoals for Employment Success** *(not included in analyses in this report)* in Baltimore and Houston. MyGoals is a coaching demonstration project designed by MDRC and partners that provides employment coaching and incentives to unemployed adults receiving housing assistance. It is operated within the Housing Authority of Baltimore City and the Houston Housing Authority, respectively.

For additional information about the evaluation and for snapshots of each program, visit https://www.acf.hhs.gov/opre/research/project/evaluation-of-coaching-focused-interventions-for-hard-to-employ-tanf-clients-and-other-low-income-populations.

Measures of self-regulation skills were collected at baseline before participants were randomly assigned to a study group and will be collected again at two follow-up time points via participant follow-up surveys—at 9 and 21 months after random assignment. For three of the four programs in the evaluation (see Box 1), measures of self-regulation skills were collected at baseline and will be collected at the two follow-up points. The evaluation did not collect self-regulation data at baseline for the MyGoals program, so information for that program is excluded throughout this report.[5]

In this report, we focus on the baseline measures of self-regulation collected during the enrollment phase of the Evaluation of Employment Coaching, from June 2018 through November 2019. The evaluation measured self-regulation skills at baseline for four reasons: (1) to provide a way to examine changes in participants' self-regulation skills over time; (2) to improve the precision of the estimates of the impacts of the programs on self-regulation skills; (3) to check that participants in the treatment and control groups did not differ systematically from each other before the start of the evaluation; and (4) to create subgroups defined by the level of self-regulation skills.

Even though all the programs target low-income individuals, the characteristics of the study participants differed in important ways across the programs.

2. Characteristics of the study participants

Even though all the programs target low-income individuals, the characteristics of the study participants differed in important ways across the programs (Table 1). For example, compared with the other two programs, LIFT served a higher proportion of Hispanic participants and participants who were older on average. In addition, 38 percent of participants in the LIFT study did not have a high school or General Educational Development (GED) diploma, whereas the corresponding figure in Goal4 It! was 22 percent. Programs also differed in the percentage of study participants who were employed in the month before applying for the programs, with employment rates ranging from 27 percent for Goal4 It! participants to 52 percent for LIFT participants. We also considered participants' perceptions about the lack of specific employment challenges. For example, overall, 27 percent reported that a lack of childcare did not make it hard for them to find or keep a good job. Participants' perceptions about these challenges also differed across programs. For example, 20 percent of LIFT participants reported that childcare was not a challenge compared to 36 percent of those from FaDSS. Along those same lines, 90 percent of LIFT participants reported that they had stable housing compared to 67 to 76 percent of participants in other programs. The variation in the programs enables us to test the properties of self-regulation measures across low-income populations with different characteristics.

[5] Study enrollment and baseline data collection were conducted using a different software platform for MyGoals than for the other programs participating in the Evaluation of Employment Coaching. The MyGoals baseline data collection did not include information on self-regulation.

Table 1. Characteristics of study participants across study programs at baseline

	Percentage of sample with each characteristic by program			
Characteristic	**All**	**FaDSS**	**LIFT**	**Goal4 It!**
Demographic and educational background				
Female (percentage)	93	94	95	90
More than 30 years old (percentage)	51	38	63	54
Race and ethnicity (percentage)				
Hispanic	41	12	71	42
Black	25	36	27	8
White	31	49	1	47
Other	2	3	1	3
Education (percentage)				
Less than high school diploma or GED	28	24	38	22
High school diploma or GED	33	41	25	32
Attended college or more	39	35	37	46
Recent employment outcomes				
Hardly ever or never having been late for a job or a job-related appointment like an interview or meeting with a program worker (percentage of those with appointments)	54	63	53	45
Worked for pay in past 30 days (percentage)	38	34	52	27
Earnings in past 30 days ($)	329	162	627	188
Employment facilitators				
Perceived lack of employment challenges related to (percentage)[a]:				
Access to transportation	33	36	36	27
Access to childcare	27	36	20	25
Access to clothes or tools for work	41	50	41	31
Adequate skills or education	32	43	26	27
Availability of jobs in area	30	36	21	34
Lack of criminal record	75	73	88	64
Lack of health condition	56	61	61	44
Possession of a valid driver's license (percentage)	52	54	42	62
Access to stable housing (percentage)	76	72	90	67
Sample size	**2,473**	**863**	**808**	**802**

Source: Evaluation of Employment Coaching baseline survey.

[a] The percentage reporting that the potential challenge made it "not at all hard" to find or keep a job.

C. STUDY MEASURE SELECTION

We selected and developed self-regulation measures that met four criteria, including that the measures:

1. **Capture skills that could be influenced by the program and could affect employment outcomes**. Coaching is hypothesized to improve employment outcomes by strengthening and fostering the use of self-regulation skills. To test whether the programs are working as hypothesized, we selected measures that capture self-regulation skills aligned with the goals of the programs and that could affect the targeted employment outcomes. Given that the links between measures of self-regulation skills and employment outcomes have not been well studied, we drew on each program's logic model.
2. **Account for other factors that could affect the self-regulation measures**. Measures of self-regulation skills are based on observed behaviors that could depend on factors other than a person's self-regulation skills. For example, a question from the Behavior Rating Inventory of Executive Function–Adult Version (BRIEF-A) asks respondents about the extent to which they "make careless errors when completing tasks." All else equal, people with higher self-regulation skills report making fewer errors. However, other factors could influence responses to this question, such as a person's unmet needs, which can distract people from focusing on a task (Mullainathan and Shafir 2014). For instance, individuals who lack financial resources might worry about paying rent, which distracts them from their work and causes them to make more errors than they might otherwise make. In this case, the responses to the question might reflect not only the individuals' self-regulation skills but also their financial resources. To help account for such issues, we selected some measures that explicitly account for other factors, including the context in which participants use self-regulation skills (for example, at work). A person's background is also an important consideration for measuring skills because most surveys were not developed for use with low-income populations and might not be phrased appropriately. For instance, depending on their level of education, respondents may not be familiar with words that appear in existing surveys.
3. **Are feasible to administer in an impact evaluation**. Long surveys or performance tasks that require access to computers or in-person administration may be infeasible for some evaluations. Short, self-reported surveys are relatively easy and quick to administer (Duckworth and Yeager 2015). To minimize burden and allow the evaluation to collect other data, the baseline and follow-up surveys allotted only about five to ten minutes to collect data on self-regulation skills. Given these constraints, we focused on self-reported measures that were administered online and by telephone. However, as discussed in Kautz and Moore (2018), other modes—such as observer reports and behaviors collected through administrative data—are promising approaches that could be used in other evaluations, if feasible.
4. **Are likely to be representative, reliable, and valid for the participants in the study**. As discussed further in Section III, measures of self-regulation skills might not represent the participants in the study if response rates are low or if responses do not vary much across study participants. By selecting measures that

are understood by respondents and pre-testing the measures, researchers can help ensure higher response rates and more representative measures. Reliability and validity relate to the extent to which measures consistently capture what they were designed to measure. Although few self-regulation measures have been validated specifically for the purpose of evaluating employment programs targeted to low-income populations, past evidence on reliability and validity can still be informative in this setting. For that reason, we prioritized measures supported by evidence of reliability and validity.

D. DEFINITIONS OF MEASURES USED

Per the criteria outlined above, we selected four primary self-regulation measures for inclusion in the baseline survey for the evaluation: Self-Esteem; Emotional Control & Self-Monitoring; Goal-Setting; and Task Monitoring, Planning, & Initiation (Table 2). These four measures spanned the three categories of self-regulation skills targeted by the programs in the evaluation: personality factors, emotional skills, and cognitive skills. The four measures comprise 26 survey items.

To minimize the burden on participants and to help to ensure high response rates, the baseline survey included self-reported measures that respondents could complete quickly (Criteria #3 and #4, above). We used existing, validated instruments when possible to help ensure that the measures would be reliable and valid for our sample (Criterion #4). Although three of the four measures were based on existing assessments, we selected three or four questions from each assessment to reduce the required response time (see Section III.B for additional guidance on selecting the number of questions). The Emotional Control & Self-Monitoring and the Task Monitoring, Planning, & Initiation measures may be divided into five total subscales.[6] To allow us to analyze the subscales separately, we selected four questions from each subscale. Per the outcome of our validation analyses, however, we focus on the overall Emotional Control & Self-Monitoring and Task Monitoring, Planning, & Initiation measures (Section III.C). Because no existing measure captured employment goal-setting—a key skill targeted by the employment programs—we developed a new three-item measure on that topic.

The measures are designed to capture a range of relevant skills and account for other factors that could influence them. Each measure has either an empirical or theoretical relationship to employment outcomes and could be influenced by the employment programs in the evaluation, as suggested by the programs' logic models (Criteria #1). The measures include both generalized measures that apply across all contexts (for example, a general measure of Self-Esteem) and contextualized measures that apply to the employment setting (for example, Goal-Setting) and therefore account for how a respondent's setting could influence the measures (Criterion #2). Compared to generalized measures, contextualized measures might be more likely to be affected by an employment program and more directly linked to specific employment outcomes. In contrast, generalized measures might be less sensitive to an employment program but apply to a broader set of outcomes and therefore might confer greater benefits in more aspects of the participant's life.

[6] We adopt the convention of capitalizing specific psychological measures.

Pre-testing is particularly important when measuring self-regulation skills among low-income populations because many measures were not developed with this population in mind.

To help ensure that study participants understood the survey questions (Criteria #2), we pre-tested the baseline and follow-up surveys. The baseline and follow-up surveys were administered in-person to eight and nine people, respectively, who were receiving services from programs for low-income people.[7] After the administration of the surveys, the interviewers conducted cognitive interviews with the respondents to ensure that they understood the questions. As a result of the pre-tests, we modified some of the questions. For example, one respondent indicated that he did not understand the word "prioritizing" when asked if he had trouble prioritizing activities. We rephrased the question to ask if respondents have trouble "deciding which activities to get done first." Pre-testing is particularly important when measuring self-regulation skills among low-income populations because many measures were not developed with this population in mind.

[7] These programs included a food bank in New Jersey, an anti-poverty program in New Jersey, and a workforce center in Colorado.

Table 2. Definitions of skills and survey measures

Skill	Definition	Relationship to employment outcomes	Measure and survey items
Personality factors			
Self-Esteem (generalized)	• Favorable attitude toward oneself	Correlated with earnings and employment (Heckman and Kautz 2012)	Rosenberg's Self-Esteem measure (Rosenberg 1965) • I am able to do things as well as most people. • I certainly feel useless at times. • All in all, I tend to feel that I am a failure. Response options: (0) strongly disagree, (1) disagree, (2) agree, (3) strongly agree
Emotional skills			
Emotional Control & Self-Monitoring (generalized)	• Emotional Control: Modulate emotional responses appropriately • Self-Monitoring: Keep track of the effect of own behavior on others; attend to own behavior in the social context	Correlation between emotional stability and job performance (r = 0.08) based on a meta-analysis of 116 studies (Barrick and Mount 1991)	Behavior Rating Inventory of Executive Function–Adult Version (Roth et al. 2005) • Eight items (redacted due to copyright) Response options: (0) never, (1) sometimes, (2) often
Cognitive skills			
Goal-Setting (contextualized)	• Set realistic employment goals	Theoretical basis that goal-setting is important as a means of developing other self-regulation skills and of attaining and maintaining employment (Babcock 2014; Locke and Latham 1990; Zimmerman et al. 1992; Cavadel et al. 2016)	New study-developed items • I set long-term employment goals that I hope to achieve within a year, such as finding a job, finding a better job, getting promoted, or enrolling in further education. • I set specific short-term goals that will allow me to achieve my long-term employment goals. • I know I need to get a job or a better job and really think I should work on finding one. Response options: (0) strongly disagree, (1) disagree, (2) agree, (3) strongly agree
Task Monitoring, Planning, & Initiation (generalized)	• Task Monitoring: Check work; assess performance during or after finishing a task to ensure attainment of a goal • Planning: Anticipate future events; set goals; develop appropriate steps to carry out an associated action; carry out tasks in a systematic manner; understand main ideas • Initiation: Begin a task or activity; fluidly generate ideas	Correlation between related measures and job performance (r = 0.22) based on a meta-analysis of 116 studies (Barrick and Mount 1991)	Behavior Rating Inventory of Executive Function–Adult Version (Roth, Isquith, and Gioia 2005) • Twelve items redacted due to copyright. Response options: (0) never, (1) sometimes, (2) often

III. Testing Measures of Self-Regulation Skills

To help ensure that the measures of self-regulation skills were appropriate for use in the evaluation's impact analyses, we tested their performance. First, we assessed the response rates and response patterns, which suggested whether the measures of self-regulation skills were representative of the study participants and whether the questions captured the full range of each self-regulation skill.

Second, we assessed reliability and validity, which provided evidence that measures captured distinct skills related to employment outcomes and that the evaluation can detect impacts on the skills. *Reliability* refers to the extent that an assessment tool produces consistent results. In this report, we focus on one form of reliability, internal consistency, which refers to the degree to which different items for a given measure produce similar results—for example, the extent to which the three Self-Esteem questions measure the same construct. *Validity* refers to the extent to which a measure captures what it is designed to measure. A measure could have a high level of internal consistency—because all the items are highly related—but could have a low level of validity because the group of items does not capture the intended construct. Although many measures of self-regulation skills have been shown to be reliable and valid, few have been tested specifically for use with low-income populations. This gap is potentially problematic because the reliability and validity of self-regulation measures can differ for different populations (Schmitt et al. 2007). In addition, the appropriate criteria for evaluating reliability and validity differ for evaluations as compared to their application for other uses, such as for diagnostic purposes in clinical settings.

To test the performance of the self-regulation measures in this study, we examined individual-level nonresponse, item-level response patterns, measures of reliability, and the results from a confirmatory factor analysis (CFA) (Box 2). We compared the performance of the measures relative to standard benchmarks.

Box 2. Description of confirmatory factor analysis (CFA) and overview of the approach used in this study

A CFA is appropriate for the validation of existing measures or those based on theory, as in the measures in this study (Brown 2015). In a CFA, the researcher first defines a model that imposes assumptions about the relationships between items in the measures. Second, the researcher conducts analyses to see whether the data support the assumptions.

In this study, we started by assuming that groups of items in the same self-regulation measure captured the same underlying skill ("factor") and that items from different measures captured different underlying skills. For example, we assumed that two items from the Self-Esteem measure captured the same skill, whereas items from the Goal-Setting measure and the Emotional Control & Self-Monitoring measure captured different skills. In particular, we specified a single model that included all 26 items and four factors that corresponded to the four self-regulation skills. We constrained the model so that each item was related to only the skill that the item was designed to measure. If the model fits the data well, then it supports this grouping of items. If not, then the model could suggest a different grouping of items.

To estimate the confirmatory factor model, we used the weighted least squares with mean and variance adjustment (robust) estimator (WLSMV), which has been shown to be both robust and feasible for models with categorical measures (Brown 2015). The survey responses were modeled as categorical variables, rather than as continuous ones, by using a probit link function.

For each analysis, we discuss the rationale for conducting the analysis, the criteria that we used to evaluate performance of the measure, and the supporting evidence for whether the measure met the criteria. To provide evidence on how much the results might change across contexts, we conducted all analyses both using the full study sample and subsamples based on each of the three employment coaching programs. In Appendix A, we report results from analyses based on different demographic groups.

The usefulness of our self-regulation measures hinges on collecting data from a set of respondents who represent the full sample of study participants.

A. TO WHAT EXTENT DO RESPONDENTS COMPLETE ITEMS?

The usefulness of our self-regulation measures hinges on collecting data from a set of respondents who represent the full sample of study participants. If only certain types of participants respond to the survey questions, then our analyses could suffer from nonresponse bias and misrepresent the participants' self-regulation skills. For example, such bias could arise if the types of participants who did not respond to the survey questions also have lower levels of self-regulation skills. For the evaluation's baseline survey, all participants took the survey as part of intake into the study. Thus, failure to respond to relevant survey items was the only source of nonresponse. We nevertheless conducted analyses to confirm that the response rates were acceptable. We also considered how participants responded to the questions. If almost all participants selected the same response to a question, then the question would not help distinguish among skill levels across respondents.

1. Definitions and rationale: Two types of response to consider

We examined two types of response patterns as follows:

a. **Survey response rates**. The overall response rate on a survey question is the percentage of study participants who complete the question. Following the National Center for Education Statistics (NCES) guidelines, we considered the overall response rate to be adequate if it exceeded 85 percent (NCES 2003). The overall response rate depends on (1) the survey response rate, or the percentage of total study participants who took the survey, and (2) the item response rate, or the percentage of participants who completed an item out of those who took the survey. We considered the response rates for each item separately because they can be informative on their own. If many respondents skip an item, the item might not perform well and could lead to nonresponse bias. For example, respondents might skip an item that is poorly phrased or difficult to understand. We considered the item response rate to be adequate if it exceeded 90 percent.

b. **Item response variance**. Item response variance refers to the extent to which different respondents mark different response categories on a given item. A large percentage of respondents providing the same answer to a question suggests that the item does not adequately distinguish among different respondents. Following NCES (2015), we considered an item to have adequate response variance if fewer than 90 percent of responses fell in each of the response categories.

2. Results: Selected measures had adequate overall response and item response variance

All items in the selected measures exhibited adequate levels of response rates and response variance (Table 3).[8] Given that completion of the baseline survey was a requirement for participation in the study, the survey response rate was 100 percent. Across all measures, the lowest level of response for a single item was 97 percent, which exceeded the target of 90 percent. The results suggest that there is no need to account for nonresponse bias when analyzing the study's baseline data. Across all measures, the greatest percentage of responses in a single category was 70 percent, which was below the cutoff of 90 percent.

Table 3. Item response patterns by self-regulation skill

	Range of item response rates (percentage)		Range of maximum percentage of responses in a single category		
Skill	**Min**	**Max**	**Min**	**Max**	**Meets criteria[a]**
Goal-Setting[b]	97	98	51	53	**Yes**
Self-Esteem[c]	97	98	38	51	**Yes**
Emotional Control & Self-Monitoring[d]	97	98	44	67	**Yes**
Task Monitoring, Planning, & Initiation[d]	97	98	45	70	**Yes**

Source: Evaluation of Employment Coaching baseline survey.

[a] The The criteria are that each measure has (1) at least a 90 percent response rate for all items and (2) fewer than 90 percent of responses in a single response category for each item.

[b] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Goal-Setting skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[c] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Self-Esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[d] A 0- to 2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

3. Considerations for other studies

Although the measures of self-regulation in this study met the criteria for adequate response rates and response variance, other studies might be more susceptible to these issues (see Box 3 for some ways to prevent and address low response rates and response variance). For example, overall response rates could be lower for studies that have a gap between identifying the sample and conducting the survey, because it could be difficult to locate the respondents. In addition, surveys could suffer from low item response if they asked about more sensitive topics.

[8] Information on response rates and response variance for each item appears in Table A.1 in Appendix A.

Box 3. Considerations for preventing and addressing low response rates and response variance

- **Preventing low survey and item response rates**. To increase survey response rates, studies may (1) collect information that can be used to follow up with nonrespondents, such as various forms of contact information for the respondent as well as information on people known by the respondent; (2) offer several survey formats that could appeal to participants (for example, web-based, telephone, and paper-and-pencil options); and (3) offer incentives for completion of surveys. To improve item response rates, researchers can pre-test the survey and modify items that respondents skip because they are confusing or too long.
- **Preventing low response variance**. Pre-testing a survey allows researchers to identify items with low response variance. For items with low response variance, researchers can modify them to increase response variance. For multiple-choice items, adjusting the response options can improve response variance by giving respondents more options that are "nearer" to the response option that was most frequently selected. The response options should still be meaningfully different from each other.
- **Addressing low overall response rates**. Researchers may test for nonresponse bias by comparing the characteristics of respondents with those of nonrespondents. If nonresponse bias is a concern, nonresponse weights can help ensure that the analyses represent the composition of the population of interest.

B. DO ITEMS DESIGNED TO MEASURE THE SAME SKILL RELATE TO EACH OTHER?

1. Definitions and rationale: Reliability is a major consideration for statistical power

Reliability. Reliability refers to the degree to which an assessment tool produces consistent results. We focus on one type of reliability—internal consistency—which is the degree to which different items for a given measure produce similar results. Internal consistency also relates to measurement error—the difference between the measure of a self-regulation skill and the true self-regulation skill. Measurement error arises when the responses to survey questions partly capture the skill of interest but also reflect other factors unrelated to the skill. For example, measurement error could arise if responses to a question depended on other factors, such as a respondent's background, rather than on the skill of interest. It could also arise if a respondent inadvertently marked the wrong response when completing the survey. Under standard assumptions, a measure with higher internal consistency also has lower levels of measurement error.[9]

We assessed internal consistency by using Cronbach's alpha, a commonly used statistic that ranges from 0 to 1; a higher value indicates that a group of items is more rather than less internally consistent. Cronbach's alpha depends on (1) the average correlation among items within a measure (Cronbach's alpha is higher when the correlation is higher) and (2) the number of items within a measure (Cronbach's alpha is higher when there are more items). Under standard assumptions, Cronbach's alpha is the proportion of the variability in a measure that depends on the true skill as opposed to measurement error (see Appendix B).

[9] In this report, we focus on classical measurement error that is assumed to be uncorrelated across items and uncorrelated with the self-regulation skills.

The criteria for assessing reliability differ for use in impact evaluations compared to use in other applications.

The criteria for assessing reliability differ for use in impact evaluations compared to use in other applications. High reliability is essential when high-stakes decisions depend on the scores of an individual. For example, reliability is especially important for a standardized test used for college admissions or for use in clinical settings. Given that impact evaluations focus on comparisons of *groups* of individuals, low reliability produces more limited consequences and does not generally lead to biased impact estimates (see Box 4 for reasons for this finding).

Box 4. Reasons why measurement error can be less consequential in an impact evaluation compared to other settings

1. **Measurement error plays less of a role when focusing on a group of individuals**. Evaluations estimate impacts by comparing the average outcomes between groups of individuals (the treatment and comparison groups); such a comparison requires an accurate estimate of, in this case, the self-regulation skills for each group rather than for each individual. In evaluations, the effect of an individual's measurement error is reduced because the errors within a group "average out" across the individuals.
2. **Estimated standard errors and statistical significance account for measurement error**. The standard error of an impact estimate is a measure of the variability of an impact estimate and affects whether an impact estimate is deemed statistically significant. A higher standard error indicates more uncertainty about the true value of the impact estimate. For a given estimate, a higher standard error reduces the statistical significance of the estimate. Higher levels of measurement error in an outcome translate into higher standard errors in the impact estimate. However, standard methods for calculating standard errors—such as those used when estimating ordinary least squares regressions—account for measurement error. Therefore, the reporting of impact estimates often accounts for a measure's internal consistency.
3. **Measurement error in outcome variables does not lead to systematic bias in impact estimates**. When self-regulation measures are analyzed as outcomes, measurement error does not lead to a systematic bias in the impact estimate. In other words, on average, the estimated impact estimate equals the true impact estimate. This finding arises because measurement error for outcomes is expected to be the same, on average, in the treatment and comparison groups. Since measurement error does not systematically affect one group more than the other, the difference between the outcomes reflects the true impact on average.
4. **Measurement error in control variables does not lead to systematic bias for some types of evaluations**. Baseline measures of self-regulation may also be used as control variables for the analysis of data from RCTs. Controlling for such measures can improve statistical precision when estimating a program's impact on outcomes. Measurement error in the self-regulation measure tends to reduce its estimated association with the outcome variable. However, if the baseline measures of self-regulation are not associated with an individual's assignment to the treatment or control group, measurement error does not bias the impact estimate. In RCTs, randomization ensures that the baseline variables are not, on average, correlated with treatment status. In other types of evaluations—such as quasi-experimental designs that rely on comparing the outcomes of treatment and comparison groups that are observationally equivalent in terms of baseline variables—measurement error in the baseline variables could lead to bias in the estimated impacts. Therefore, ensuring a higher reliability is more important for such designs.

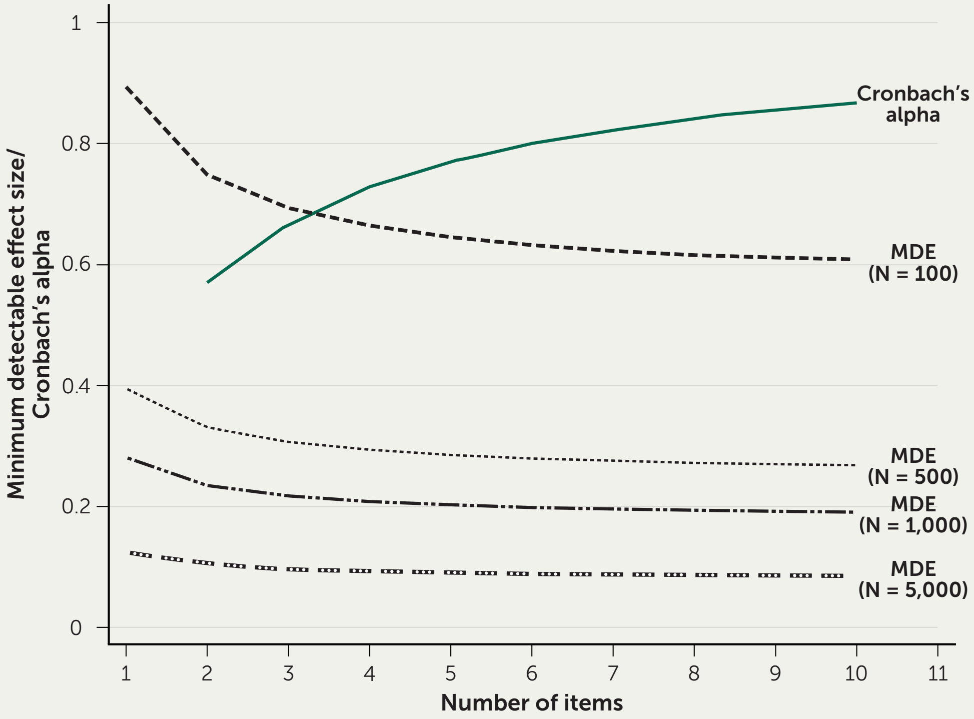
Because low reliability has limited consequences in impact evaluations, we recommend an examination of internal consistency, but we endorse standard thresholds for acceptable levels of internal consistency as rules of thumb rather than as hard-and-fast rules. Despite debate in the psychological literature about an appropriate threshold for an acceptable Cronbach's alpha, some consensus suggests that a value in the range of 0.65

to 0.70 represents a minimum threshold (DeVellis 2017; Bland and Altman 1997). We suggest aiming for a Cronbach's alpha within this range but advise against dismissing measures with slightly lower values.

For the selection of self-regulation measures, we recommend consideration of how the reliability and number of items will affect statistical power. In an impact evaluation, a primary reason to consider reliability is that measurement error can affect statistical power, which is often summarized by the minimum detectable effect (MDE)—defined as the smallest impact in standard deviation units that the study could estimate as statistically significant. Studies can detect smaller effects when measures involve less measurement error because measures with less error exhibit less unexplained variance, making it easier to attribute any observed differences in study groups to the program rather than to the variance in the measure. Given that self-regulation measures with more items have less measurement error for a given average inter-item correlation, the inclusion of more items per measure improves the capacity to detect smaller effects for the measure. Even though the addition of items improves reliability and thus statistical power, it also increases the burden on study participants as well as study costs.

When selecting self-regulation measures for this study, we considered the tradeoff between burden and statistical power. In Figure 1, we provide a sense of this tradeoff by displaying the Cronbach's alpha and MDEs as a function of the number of items in

Figure 1. Cronbach's alpha and minimum detectable effect sizes as a function of the number of items in a measure



Notes: The MDEs were calculated by using a standard formula for randomized controlled trials with individual-level random assignment (Schochet 2008) but were adapted to include measurement error (see Appendix B for a description of this adaptation). We assumed that the inter-item correlation is 0.40 between all pairs of items, measurement error is uncorrelated with the true skill, and measurement error is uncorrelated across items.

a measure for hypothetical RCTs of varying sample sizes. Consistent with Figure 1, we selected three or four items for each measure.[10] With our sample size of over 800 participants per program, the addition of more items would have led to minimal improvements in statistical power. However, a study with a sample of 100 participants could improve statistical power by including more than four items. This example illustrates that Cronbach's alpha is not enough to summarize whether a measure is sufficiently reliable to meet an evaluation's needs, because the ideal level of reliability could also depend on the sample size. Researchers planning an evaluation can calculate statistical power for various numbers of items if they have access to an estimate of the Cronbach's alpha for an existing measure (see Appendix B for additional details).

Standardized factor loadings. Our confirmatory factor analysis (Box 2) provides a complementary statistic—called the factor loading—that captures the extent to which each item relates to the corresponding self-regulation skill. A factor loading ranges from -1 to 1, and a positive (negative) factor loading indicates that higher values on the item are positively (negatively) associated with the overall measure. The sign of the factor loading (negative or positive) should match the expected relationship with the skill. For example, the item "I certainly feel useless at times" should have a negative loading on Self-Esteem, whereas the item "I am able to do things as well as most people" should have a positive loading. If the absolute value of a factor loading is low, then the item might not fit with a particular self-regulation skill. Removing items with a low factor loading can improve the reliability of the measure. The literature suggests the use of 0.40 as a threshold for "low" (Stevens 2012).

Our results suggest that the self-regulation measures exhibit sufficient reliability for use in impact evaluations.

2. Results: The items designed to measure a given skill consistently measure that skill (are sufficiently reliable)

Our results suggest that the self-regulation measures exhibit sufficient reliability for use in impact evaluations. In the full sample, the estimated Cronbach's alpha is 0.65 or above for all scales, meeting our rule-of-thumb threshold for reliability (Table 4). Similarly, the factor loadings are all greater than 0.40 for the full sample, suggesting that each item fits well with the other items in the self-regulation measure (Table 5). Compared to the Self-Esteem measure and the Goal-Setting measure, the Emotional Control & Self-Monitoring and the Task Monitoring, Planning, & Initiation measures demonstrate higher reliability. The difference could have arisen either because the latter two measures include more items than do the other measures or because they have higher inter-item correlations, or both. In Table 4, we show that all measures have similar inter-item correlations, indicating that the higher reliabilities were primarily attributable to the inclusion of more items. The reliability varies across samples, with the LIFT program's sample members tending to exhibit the lowest reliabilities. Similar patterns arise across samples defined by other demographic characteristics (Tables A.2 and A.3 in Appendix A).

[10] As discussed in Section II, we selected four items for each of the subscales within the Emotional Control & Self-Monitoring measure and the Task Monitoring, Planning, & Initiation measures.

Table 4. Reliability of measures of self-regulation skills across samples

	Cronbach's alpha (α) and inter-item correlation (r) by self-regulation skill							
	Goal-Setting[a] (3 items)		Self-Esteem[b] (3 items)		Emotional Control & Self-Monitoring[c] (8 items)		Task Monitoring, Planning, & Initiation[c] (12 items)	
Sample	α	r	α	r	α	r	α	r
All	**0.65**	**0.39**	**0.65**	**0.37**	**0.85**	**0.42**	**0.88**	**0.38**
Program								
FaDSS	0.69	0.43	0.66	0.38	0.85	0.42	0.87	0.36
Goal4 It!	0.68	0.42	0.65	0.37	0.85	0.42	0.91	0.45
LIFT	0.56	0.30	0.62	0.34	0.83	0.39	0.87	0.36

Source: Evaluation of Employment Coaching baseline survey.

[a] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Goal-Setting skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[b] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Self-Esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[c] A 0- to 2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

Table 5. Range of factor loadings of self-regulation items across samples

	Range of factor loadings by self-regulation skill							
	Goal-Setting[a]		Self-Esteem[b]		Emotional Control & Self-Monitoring[c]		Task Monitoring, Planning, & Initiation[c]	
Sample	Min	Max	Min	Max	Min	Max	Min	Max
All	**0.42**	**0.84**	**0.50**	**0.85**	**0.69**	**0.81**	**0.58**	**0.79**
Program								
FaDSS	0.47	0.85	0.52	0.84	0.70	0.80	0.57	0.78
Goal4 It!	0.46	0.83	0.50	0.86	0.69	0.87	0.64	0.87
LIFT	0.32	0.80	0.44	0.86	0.65	0.81	0.55	0.78

Source: Evaluation of Employment Coaching baseline survey.

Notes: The estimates for each sample come from a single confirmatory factor model that assumes four factors that correspond to the four self-regulation skills. The items corresponding to each skill are constrained to relate only to that skill. The factors are not constrained to be independent.

[a] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Goal-Setting skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[b] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Self-Esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[c] A 0- to 2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

3. Considerations for other studies

Our findings are consistent with those from other studies that have tested similar measures. Using two samples, the initial psychometric validation of the BRIEF-A found Cronbach's alphas between 0.93 and 0.97 for the Emotional Control & Self-Monitoring and the Task Monitoring, Planning, & Initiation measures (Roth et al. 2005). However, we used versions with fewer items (eight versus 30 for Emotional Control and Self-Monitoring and 12 versus 40 for Task Monitoring, Planning, & Initiation). If we recalculated Cronbach's alpha by using the number of items from the original measures, but with the inter-item correlations from our sample, the values would be 0.96 and 0.96 for both measures, much closer to the original estimates. Similarly, earlier studies of the ten-item version of the Self-Esteem measure found a higher reliability of 0.77 (Rosenberg 1979). If we recalculate the reliability of the three-item version but assume ten items, the reliability is 0.85.

These findings suggest that our measures had lower reliabilities primarily because they included fewer items, not because the items did not reflect the underlying self-regulation skills. Therefore, the main reason to include more items in this study would have been to improve statistical power. However, given the study's sample size, an increase in the number of items would have had a marginal impact on statistical power (Figure 1). If the reliability of the measures had not been acceptable, we could have taken several steps to address the low reliability (Box 5).

Box 5. Considerations for studies with measures that do not meet criteria for adequate reliability

1. **Removing items with low factor loadings**. Sometimes items with low factor loadings can reduce the overall reliability of a measure. One option is to remove such items from the measure to increase reliability. Removing an item, however, does not always improve reliability. On one hand, removing the item with a low factor loading could increase the average inter-item correlation, which tends to increase reliability. On the other hand, removing an item means that there are fewer items in the measure, which tends to reduce reliability, keeping all else equal.

2. **Rewriting items with low factor loadings**. Sometimes individual items are not highly correlated with the other items in a measure because they capture a slightly different construct. Rewriting the items can address this issue for future survey administrations.

3. **Expanding the number of items per skill**. Because reliability depends on the number of items in a measure, an increase in the number of items tends to improve reliability in future survey administrations, if the new items measure the same construct.

4. **Reporting impacts on the individual items rather than on the overall measure**. The three approaches listed above might not help improve reliability or might be infeasible because they require additional survey administrations. An alternative is to report impacts on each item separately in addition to impacts on the overall measure. This approach offers advantages and disadvantages. One advantage is that it accounts for the possibility that the measure had low reliability because the items captured meaningful but different constructs. For example, one item might capture a slightly different aspect of a self-regulation skill that is not related to the other items in a measure. One disadvantage is that this approach introduces more total outcomes to the analysis, potentially increasing the need to correct for several hypotheses to avoid finding statistically significant results by chance.

C. DO DIFFERENT SELF-REGULATION MEASURES CAPTURE DISTINCT SKILLS?

1. Definitions and rationale: Correlations can summarize whether different measures capture distinct skills

Another result of the CFA is the estimated correlation between each pair of skills. The correlations shed light on discriminant validity, which is the extent to which separate measures capture different constructs. If two measures are very highly correlated, then they likely capture the same underlying construct; either one of them is redundant, or the theoretical basis of the model is incorrect. Sometimes measures demonstrate a high correlation even if they capture conceptually different constructs. For example, it is possible that the Task Monitoring, Planning, & Initiation and the Goal-Setting measures could be highly correlated if people set goals to improve skills in the areas of Task Monitoring, Planning, & Initiation. Therefore, theoretical considerations may also determine if measures capture distinct concepts. In the context of an impact evaluation, reliance on two measures that capture the same construct can be misleading because a program's impact will necessarily be similar for each measure. We considered two measures to have sufficient discriminant validity if the correlation between them is less than 0.80 and they are theoretically distinct (Brown 2015).

2. Results: The self-regulation measures capture distinct skills (exhibit discriminant validity)

The correlations among all the factors met our criteria for discriminant validity for all samples (correlation less than 0.80), suggesting that the measures capture distinct self-regulation skills (Table 6). Across the skills and samples, Emotional Control & Self-Monitoring and Task Monitoring, Planning, & Initiation skills are consistently the most highly correlated. This finding is not surprising given that these two measures both come from the BRIEF-A, which measured executive functioning, a broader concept about the ability to control emotions and behavior.

We also considered the correlations between the subscales of the measures of the BRIEF-A. The two measures from the BRIEF-A may be disaggregated into five subscales: (Emotional Control, Self-Monitoring, Tasking Monitoring, Planning, and Initiation). As discussed in Section II, we included a sufficient number of items to analyze the subscales separately. When we conducted the analyses with the five subscales in place of the two overall measures, we found that the correlation between skills was 0.74 for the subscales comprising Emotional Control & Self-Monitoring and 0.87, 0.89, and 0.93 for the subscales comprising Tasking Monitoring, Planning, & Initiation. Because of the relatively high correlations and conceptual similarity of the subscales, we focused on the two overall measures. [11]

[11] To help ensure that the Emotional Control & Self-Monitoring and Task Monitoring, Planning, & Initiation measures each captured a single skill, we conducted a test of the number of skills captured by each set of items. Applying the Kaiser criterion with Horn's adjustment for sampling error, we found evidence that the Emotional Control & Self-Monitoring and Task Monitoring, Planning, & Initiation measures each captured a single skill (Horn 1965; Kaiser 1960).

Table 6. Correlations between measures of self-regulation skills overall and across programs

		Correlations by sample			
Skill 1	Skill 2	All	FaDSS	LIFT	Goal4 It!
Goal-Setting[a]	Self-Esteem[b]	0.32	0.33	0.25	0.36
Goal-Setting[a]	Emotional Control & Self-Monitoring[c]	0.19	0.22	0.07	0.25
Goal-Setting[a]	Task Monitoring, Planning, & Initiation[c]	0.27	0.25	0.25	0.30
Self-Esteem[b]	Emotional Control & Self-Monitoring[c]	0.57	0.64	0.43	0.64
Self-Esteem[b]	Task Monitoring, Planning, & Initiation[c]	0.62	0.64	0.46	0.71
Emotional Control & Self-Monitoring[c]	Task Monitoring, Planning, & Initiation[c]	0.68	0.70	0.65	0.72

Source: Evaluation of Employment Coaching baseline survey.

Notes: The estimates from each sample come from a single confirmatory factor model that assumes four factors that correspond to the four self-regulation skills. The items corresponding to each skill are constrained to load only on that skill. The factors are not constrained to be independent.

[a] 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Goal-Setting skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[b] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Self-Esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[c] A 0- to 2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

3. Considerations for other studies

Our estimate of the correlations between the two measures from the BRIEF-A (Emotional Control & Self-Monitoring and Task Monitoring, Planning, & Initiation) is similar to the estimate from the original validation study (Roth, Isquith, and Gioia 2005). The original validation study found correlations of 0.78 and 0.80, depending on the sample. We estimated correlations ranging between 0.65 and 0.72. Although the findings from our analyses suggested no changes to the measures, we would have considered adjusting the measures if we had found higher correlations. More generally, if two separate measures have correlations above 0.80 and are theoretically similar, researchers might consider combining them into a single measure. In an impact evaluation, such an approach reduces the number of outcome measures, which also helps with interpretation and reduces the need to account for several hypothesis tests.

Model fit statistics are a summary of whether a confirmatory factor model fits the data well overall (exhibits model validity).

D. ARE THE SURVEY ITEMS GROUPED WITH THE APPROPRIATE SELF-REGULATION MEASURES?

1. Definitions and rationale: Model fit statistics suggest whether survey items are grouped with the appropriate skills

Model fit statistics are a summary of whether a confirmatory factor model fits the data well overall (exhibits model validity). As described at the beginning of Section III, we estimated a single confirmatory factor model that assumed four self-regulation skills, with each item related only to the skill it was designed to measure. Therefore, in this study, fit statistics provide evidence on whether the items generally group together by the self-regulation skills they were designed to measure. For example, the model might fit poorly if some items fit better with a different self-regulation skill. We calculated three standard measures of model fit and selected criteria for acceptable fit based on the literature (Table 7).

Table 7. Overall fit statistics and criteria for acceptable fit

Statistic for overall model fit	Criterion for acceptable fit
Root mean square error of approximation (Steiger and Lind 1980)	0.05 or below for a close fit and 0.08 or below for a reasonable fit as suggested by Browne and Cudeck (1992) based on practical experience
Comparative Fit Index (Bentler 1990b)	0.90 or above as suggested by Brown (2015) based on analysis by Bentler (1990a)
Tucker Lewis Index (Tucker and Lewis 1973)	0.90 or above as suggested by Brown (2015) based on analysis by Bentler (1990a)

2. Results: The grouping of items into different self-regulation measures fits the data well overall, suggesting that the items are grouped with the appropriate measures

For all samples, the fit statistics met our criteria for an acceptable fit (Table 8). The fit changes little across programs, suggesting that the basic structure of the measures applies in different contexts. The results are also similar for samples based on the different demographic subgroups we considered (Table A.4 in Appendix A). Based on these results, we made no modifications to the measures to improve their performance.

3. Considerations for other studies

If we had found that the model did not adequately fit the data, we would have considered modifications to the model and groupings of items into measures to improve the model's fit (Box 7).

Table 8. Overall fit statistics across samples

	Fit statistics				
	Root mean square error of approximation (RMSEA)				
Sample	**Estimate**	**95 percent confidence interval (lower and upper bounds)**		**Comparative Fit Index (CFI)**	**Tucker Lewis Index (TLI)**
All	**0.06**	**0.06**	**0.06**	**0.94**	**0.93**
Program					
FaDSS	0.06	0.06	0.06	0.94	0.93
Goal4 It!	0.06	0.06	0.07	0.95	0.95
LIFT	0.05	0.05	0.05	0.95	0.94

Source: Evaluation of Employment Coaching baseline survey.

Notes: The estimates from each sample come from a single confirmatory factor model that assumes four factors corresponding to the four self-regulation skills. The items corresponding to each skill are constrained to relate only to that skill. The factors are not constrained to be independent.

Box 6. Considerations for studies with measures that do not meet criteria for adequate model fit

- **Examine modification indices**. As a first step, we suggest an examination of the "modification indices," which are statistics produced from estimating a CFA that indicate how the overall model fit would change if the model were modified in a particular way. For example, the modifications indices might indicate that the model would fit the data better if one item were switched to a different self-regulation measure. After identifying a potential modification, we suggest re-estimating the model to confirm the improved fit. We recommend making few modifications because the introduction of several changes can lead to a model that might overfit the data in the study sample, thereby reducing the generalizability of results.
- **Consider an exploratory factor analysis**. If a few changes do not sufficiently improve model fit, researchers could consider an exploratory factor analysis (EFA), which places fewer restrictions on the data (in other words, it does not assume that items group together by topic). For example, in an EFA, a single item could relate to several self-regulation skills. Nevertheless, it still requires specification of the number of factors represented by the items.

E. DO THE SELF-REGULATION MEASURES RELATE TO OTHER KEY VARIABLES?

1. Definitions and rationale: Correlations between measures of self-regulation skills and employment variables provide evidence on the real-world importance of the skills

Our ability to study whether the employment programs affect later outcomes through self-regulation skills hinges on whether self-regulation skills are related to employment outcomes. Therefore, it is helpful to ensure that the self-regulation measures capture the intended construct and reflect "real-world" variables. If self-regulation skills are not related to employment outcomes, then changes in self-regulation skills would not be expected to influence employment outcomes.

Our ability to study whether the employment programs affect later outcomes through self-regulation skills hinges on whether self-regulation skills are related to employment outcomes.

Ideally, we might test whether the self-regulation measures predict future employment outcomes (predictive validity). As we do not yet have data on future employment outcomes, we focus on establishing concurrent validity—whether the self-regulation measures are correlated with other measures collected at approximately the same time.[12] We validate the self-regulation measures against two types of employment-related measures collected at baseline: (1) those directly related to recent employment outcomes and (2) the respondents' perceived lack of employment challenges, as measured using questions that ask about the extent to which potential challenges make it hard to find or keep a job. We view employment challenges as a link between self-regulation skills and employment outcomes, because addressing employment challenges is often a first step to securing employment. For example, an individual might need to secure access to childcare or transportation before finding a job. We anticipate that participants with higher levels of self-regulation skills can better meet these specific challenges and secure long-term employment.

Unlike in the case of the other analyses in this report, the criteria for determining whether the measures meet concurrent validity are less firmly established. Concurrent validity is often established by estimating the correlation between scores on one psychological assessment and a different psychological assessment designed to measure a similar construct. The estimated correlations are compared to benchmarks for a "high" correlation. However, as pointed out by Heckman and Kautz (2012), this validation approach assumes that the other assessment is valid. Instead, we adopt a more direct approach by comparing self-regulation measures to employment-related measures rather than to other self-regulation assessments. We did not find a clear benchmark for a "high" correlation for our more direct approach. Instead, we based our conclusions on whether the correlations were statistically significant.

2. Results: The self-regulation measures are correlated with contemporaneously measured employment outcomes as well as with individuals' perceived lack of employment challenges (exhibit concurrent validity)

Our analyses suggest that the four measures of self-regulation skills relate to recent employment outcomes as well as to individuals' perceived lack of employment challenges (Table 9). All the reported correlations are scaled so that a higher correlation indicates a more favorable relationship between each self-regulation measure and an employment-related measure. For example, the positive relationship between Self-Esteem and lack of a health condition indicates that individuals who reported higher levels of Self-Esteem are less likely to report a health condition that poses a challenge to finding a job. Although all the self-regulation measures are correlated with the employment-related measures, Self-Esteem most consistently has a statistically significant correlation with the employment-related measures.

[12] Data on employment outcomes is forthcoming via participant follow-up surveys currently in the field. The data collection for the first follow-up survey is scheduled to continue through late 2020, with analysis of these data conducted through the end of 2021. The data collection for the second follow-up survey is scheduled to continue through late 2021, with analysis of these data conducted through the end of 2022.

Table 9. Correlations between measures of self-regulation skills and employment-related variables

	Correlation coefficients by skill			
Variable	Goal-Setting[a]	Self-Esteem[b]	Emotional Control & Self-Monitoring[c]	Task Monitoring, Planning, & Initiation[c]
Recent employment outcomes				
Frequency of having been on time for a job or a job-related appointment such as an interview or meeting with a program worker	0.02	0.17***	0.21***	0.27***
Worked for pay in past 30 days	0.00	0.09***	0.03	0.02
Earnings in past 30 days	-0.01	0.07***	0.04**	-0.01
Employment facilitators				
Perceived lack of employment challenges related to[d]:				
Access to transportation	0.00	0.10***	0.10***	0.15***
Access to childcare	-0.04*	0.05***	0.02	0.08***
Access to clothes or tools for work	-0.05**	0.11***	0.08***	0.14***
Adequate skills or education	-0.02	0.13***	0.08***	0.19***
Availability of jobs in area	-0.05**	0.07***	0.05**	0.14***
Lack of criminal record	-0.03	0.09***	0.13***	0.10***
Lack of health condition	0.07***	0.25***	0.21***	0.23***
Possession of a valid driver's license	-0.03	-0.02	0.01	-0.01
Access to stable housing	0.00	0.12***	0.06***	0.01

Source: Evaluation of Employment Coaching baseline survey.

[a] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Goal-Setting skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[b] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Self-Esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[c] A 0- to 3-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

[d] The extent to which respondents reported that a potential challenge made it hard to find or keep a job.

* Significantly different from zero at the .10 level, two-tailed test.

** Significantly different from zero at the .05 level, two-tailed test.

*** Significantly different from zero at the .01 level, two-tailed test.

3. Discussion

The strength of the relationship between self-regulation and employment outcomes informs the value of measuring self-regulation in evaluations of programs that aim to improve employment. Although the results of this analysis provide evidence that the self-regulation measures are associated with significant employment variables, the interpretation of the analysis is not straightforward. For example, Self-Esteem and earnings exhibit a statistically significant correlation, but the reason is not clear. Self-Esteem could directly affect earnings because more confident people could be more successful in performing in a job. However, the reverse could also be true—earning more money could increase an individual's Self-Esteem. Similar reasoning could explain why Goal-Setting is negatively correlated with individuals' perceived lack of some employment-related challenges. For example, people who face particularly difficult employment challenges might need to set more goals in order to succeed in the labor market. Follow-up data on the self-regulation measures could help illuminate some of these issues. For example, if Self-Esteem predicted future earnings—holding baseline earnings fixed—then it is more likely that Self-Esteem led to higher earnings. We will investigate this possibility as part of the Evaluation of Employment Coaching when follow-up data become available.

IV. Conclusions

> The empirical results presented here demonstrate the feasibility of developing reliable and valid measures of self-regulation skills for use in impact evaluations of employment programs for low-income populations.

Through a presentation of empirical evidence, this report has demonstrated a process for developing and testing self-regulation measures in the context of an impact evaluation of employment coaching programs for low-income populations. For three reasons, it is challenging to develop and test measures of self-regulation skills in this context. First, such evaluations have feasibility constraints and need to minimize the burden on individuals. Second, most measures of self-regulation skills were developed for other purposes. Third, the standard criteria for assessing the performance of measures do not always apply to evaluations. To meet these challenges, we recommend (1) basing self-regulation measures on existing measures as much as possible; (2) selecting a subset of items from existing measures to reduce burden; (3) using a set of both general measures of self-regulation as well as measures specific to the employment context; (4) pre-testing measures to ensure that they align with the background of respondents; and (5) assessing the reliability and validity of the measures but considering carefully whether standard criteria for acceptable levels apply.

The empirical results presented here demonstrate the feasibility of developing reliable and valid measures of self-regulation skills for use in impact evaluations of employment programs for low-income populations (Table 10). Although the four measures we selected do not cover all aspects of self-regulation, we found that they performed well for use in this evaluation. In this report, we focused on selecting and testing measures in the context of employment programs that serve low-income populations, but some of our conclusions also apply to impact evaluations of other types of programs and to those that serve different populations, especially because our results are relatively stable across different programs and subpopulations (Tables A.2 – A.4 in Appendix A). As part of the Evaluation of Employment Coaching, a future report—expected to be released in fall 2021—will present estimates of the short-term impact of the coaching programs on self-regulation skills and employment outcomes. A second future report—expected to be released in fall 2022—will present estimates of longer-term impacts.

Table 10. Summary of analyses

Research question	Summary of results
To what extent do respondents complete items?	**Both the survey as a whole and each individual item met the criteria for adequate levels of response rates**. All (100 percent) respondents took the survey. Response rates for individual items were at least 97 percent. These results suggest that there is no need to account for nonresponse bias in analyses of the baseline survey. The responses also exhibited sufficient variation. Across all measures, the greatest percentage of responses in a single category was 70 percent.
Do items designed to measure the same skill relate to each other?	**The items designed to measure a given skill consistently measured that skill (were reliable)**. In the full sample, the estimated Cronbach's alpha (a measure of reliability) was 0.65 or above for all measures, meeting our target threshold. In addition, each item had a sufficiently high factor loading, a measure of the relationship between individual items and the skills. The estimated reliability differed somewhat across employment programs.
Do different self-regulation measures capture distinct skills?	**The self-regulation measures captured distinct skills (exhibited discriminant validity)**. For all samples, the correlation between pairs of skills was 0.72 or less, which met our criteria that distinct skills should demonstrate correlations below 0.80.
Are the survey items grouped with the appropriate self-regulation measures?	**The grouping of items into different self-regulation measures fit the data well overall (exhibited model validity), suggesting that the items were grouped with the right measures**. Across all samples, estimated fit statistics met standard criteria for acceptable fit. The results were stable across employment programs, suggesting some level of generalizability.
Do the self-regulation measures relate to other important variables?	**The self-regulation measures were correlated with contemporaneously measured employment outcomes as well as with individuals' perceived lack of employment challenges (exhibited concurrent validity)**. Although these findings did not lend themselves to straightforward interpretation, the results suggested that the self-regulation measures may allow us to study whether the employment programs affect later outcomes through self-regulation skills.

References

Almlund, M., A. Duckworth, J.J. Heckman, and T. Kautz. "Personality Psychology and Economics." In *Handbook of the Economics of Education*, edited by E.A. Hanushek, S. Machin, and L. Woessmann (Vol. 4, pp. 1–181). Amsterdam: Elsevier, 2011.

Babcock, E. "Using Brain Science to Design New Pathways Out of Poverty." Boston, MA: Crittenton Women's Union, 2014.

Bentler, P.M. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin*, vol. 107, no. 2, 1990a, pp. 238–246.

Bentler, P.M. "Fit Indexes, Lagrange Multipliers, Constraint Changes and Incomplete Data in Structural Models." *Multivariate Behavioral Research*, vol. 25, no. 2, 1990b, pp. 163–172.

Bland, J.M., and D.G. Altman. "Cronbach's Alpha." *BMJ (Clinical Research Ed.)*, vol. 314, no. 7080, 1997, p. 572.

Brown, T.A. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Publications, 2015.

Browne, M.W., and R. Cudeck. "Alternative Ways of Assessing Model Fit." *Sociological Methods Research*, vol. 21, no. 2, 1992, pp. 230–258.

Cavadel, E.W., J.F. Kauff, M.A. Anderson, S. McConnell, and M. Derr. "Self-Regulation and Goal Attainment: A New Perspective for Employment Programs." OPRE Report #2017-12. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2016.

Cavadel, E.W., J.F. Kauff, A. Person, and T.K. Kravis. "Perspectives on Practice: A Guide to Measuring Self-Regulation and Goal-Related Outcomes in Employment Programs." OPRE Report #2018-37. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2018.

Cronbach, L.J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, vol. 16, no 3, 1951, pp. 297–334.

DeVellis, R.F. *Scale Development: Theory and Applications, Fourth Edition*. Los Angeles: Sage Publications, 2017.

Duckworth, A.L., and D.S. Yeager. "Measurement Matters: Assessing Personal Qualities Other than Cognitive Ability for Educational Purposes." *Educational Researcher*, vol. 44, no. 4, 2015, pp. 237–251.

Hamoudi, A., D.W. Murray, L. Sorensen, and A. Fontaine. "Self-Regulation and Toxic Stress Report 2: A Review of Ecological, Biological, and Developmental Studies of Self-Regulation and Stress." OPRE Report #2015-30. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2014.

Heckman, J.J., and T. Kautz. "Hard Evidence on Soft Skills." *Labour Economics*, vol. 19, no. 4, 2012, pp. 451–464.

Horn, J.L. "A Rationale and Test for the Number of Factors in Factor Analysis." *Psychometrika*, vol. 30, no. 2, 1965, pp. 179–185.

Kaiser, H.F. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement*, vol. 20, no. 1, 1960, pp. 141–151.

Kautz, T., and Q. Moore. "Measuring Self-Regulation Skills in Evaluations of Employment Programs for Low-Income Populations: Challenges and Recommendations." OPRE Report #2018-83. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2018.

Locke, E., and G. Latham. *A Theory of Goal Setting & Task Performance*. Upper Saddle River, NJ: Prentice Hall, 1990.

Mullainathan, S., and E. Shafir. "Freeing Up Intelligence." *Scientific American Mind*, vol. 21, no. 1, 2014, pp. 58–63.

Murray, D.W., K. Rosanbalm, C. Christopoulos, and A. Hamoudi. "Self-Regulation and Toxic Stress Report 1: Foundations for Understanding Self-Regulation from an Applied Perspective." OPRE Report #2015-21. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2014.

Murray, D.W., K. Rosanbalm, and C. Christopoulos. "Self-Regulation and Toxic Stress Report 4: Implications for Programs and Practice." OPRE Report #2016-97. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2015.

Murray, D.W., K. Rosanbalm, and C. Christopoulos. "Self-Regulation and Toxic Stress Report 3: A Comprehensive Review of Self-Regulation Interventions." OPRE Report #2016-34. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2016.

Murray, D.W., and K. Rosanbalm. "Current Gaps and Future Directions for Self-Regulation Intervention Research: A Research Brief." OPRE Report #2017-93. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2017.

Murray, D.W., K. Rosanbalm, C. Christopoulos, and A. Meyer. An Applied Contextual Model for Promoting Self-Regulation Enactment Across Development: Implications for Prevention, Public Health, and Future Research. *The Journal of Primary Prevention*, vol. 40, 2019, pp. 367-403.

NCES. "ED School Climate Surveys (EDSCLS): National Benchmark Study 2016. Appendix D. EDSCLS Pilot Test 2015 Report." Washington, DC: National Center for Education Statistics, Institute for Education Sciences, U.S. Department of Education, 2015.

NCES. "NCES Statistical Standards." NCES Report #2003-601. Washington, DC: National Center for Education Statistics, Institute for Education Sciences, U.S. Department of Education, 2003.

Nunnally, J.C., and I.H. Bernstein. *Psychometric Theory*. 3rd ed. New York: McGraw–Hill, 1994.

Rosenberg, M. *Conceiving the Self*. New York: Basic Books, 1979.

Rosenberg, M. *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press, 1965.

Roth, R.M., P.K Isquith, and G.A. Gioia. *Behavior Rating Inventory of Executive Function—Adult Version: Professional Manual*. Lutz, FL: Psychological Assessment Resources, 2005.

Schmitt, D.P., J. Allik, R.R. McCrae, V. Benet-Martínez, L. Alcalay, and L. Ault. "The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description across 56 Nations." *Journal of Cross-Cultural Psychology*, vol. 38, no. 2, 2007, pp. 173–212.

Schochet, P.Z. "Statistical Power for Random Assignment Evaluations of Education Programs." *Journal of Educational and Behavioral Statistics*, vol. 33, no. 1, 2008, pp. 62–87.

Steiger, J.H., and J.C. Lind. "Statistically Based Tests for the Number of Common Factors." Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Stevens, J.P. *Applied Multivariate Statistics for the Social Sciences*. New York: Routledge, 2012.

Stroop, J.R. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology*, vol. 18, no. 6, 1935, p. 643.

Tucker, L.R., and C. Lewis. "A Reliability Coefficient for Maximum Likelihood Factor Analysis." *Psychometrika*, vol. 38, no. 1, 1973, pp. 1–10.

Zimmerman, B., A. Bandura, and M. Martinez-Pons. "Self-Motivation for Academic Attainment: The Role of Self-Efficacy Beliefs and Personal Goal Setting." *American Educational Research Journal*, vol. 29, no. 3, fall 1992, pp. 663–667.

Appendix A: Supplementary information

Table A.1. Item response patterns by skill

Skill	Item	Response rate	Percentage of respondents who choose the category with the most responses	Meets criteria
Goal-Setting[a]				**Yes**
	1	98	53	
	2	97	53	
	3	97	51	
Self-Esteem[b]				**Yes**
	1	98	51	
	2	97	38	
	3	97	41	
Emotional Control & Self-Monitoring[c]				**Yes**
Emotional Control	1	97	45	
	2	97	44	
	3	97	67	
	4	97	55	
Self-Monitoring	1	97	54	
	2	97	47	
	3	97	64	
	4	98	66	
Task Monitoring, Planning, & Initiation[c]				**Yes**
Task Monitoring	1	98	57	
	2	97	70	
	3	98	55	
	4	97	61	
Planning	1	97	65	
	2	97	52	
	3	98	50	
	4	98	52	
Initiation	1	97	52	
	2	97	55	
	3	97	66	
	4	97	45	

Source: Evaluation of Employment Coaching baseline survey.

[a] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Goal-Setting skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[b] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Self-Esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[c] A 0- to 2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

Table A.2. Reliability of measures of self-regulation skills across samples

	Cronbach's alpha (*α*) and inter-item correlation (*r*) by self-regulation skill							
Sample	**Goal-Setting[a] (3 items)**		**Self-Esteem[b] (3 items)**		**Emotional Control & Self-Monitoring[c] (8 items)**		**Task Monitoring, Planning, & Initiation[c] (12 items)**	
	α	*r*	*α*	*r*	*α*	*r*	*α*	*r*
All	**0.65**	**0.39**	**0.65**	**0.37**	**0.85**	**0.42**	**0.88**	**0.38**
Program								
FaDSS	0.69	0.43	0.66	0.38	0.85	0.42	0.87	0.36
Goal4 It!	0.68	0.42	0.65	0.37	0.85	0.42	0.91	0.45
LIFT	0.56	0.30	0.62	0.34	0.83	0.39	0.87	0.36
Gender								
Female	0.65	0.39	0.66	0.38	0.85	0.43	0.88	0.38
Male	0.68	0.43	0.55	0.27	0.83	0.38	0.87	0.37
Age								
Less than or equal to 30 years old	0.66	0.40	0.66	0.38	0.85	0.42	0.87	0.35
More than 30 years old	0.64	0.38	0.65	0.36	0.85	0.43	0.89	0.41
Race and ethnicity								
Hispanic	0.68	0.42	0.58	0.30	0.85	0.41	0.88	0.39
Black	0.52	0.27	0.66	0.39	0.83	0.38	0.86	0.34
White, other	0.67	0.41	0.67	0.39	0.85	0.42	0.88	0.39
Education								
Less than high school diploma or GED	0.62	0.35	0.62	0.34	0.85	0.41	0.88	0.39
High school diploma or GED	0.67	0.41	0.65	0.36	0.85	0.42	0.88	0.37
College	0.66	0.40	0.68	0.40	0.86	0.43	0.88	0.39
Employment								
Worked for pay in past 30 days	0.58	0.33	0.65	0.37	0.85	0.41	0.87	0.37
Did not work for pay in past 30 days	0.70	0.44	0.66	0.37	0.86	0.43	0.89	0.39

Source: Evaluation of Employment Coaching baseline survey.

[a] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Goal-Setting skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[b] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Self-Esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[c] A 0- to 2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

Table A.3. Range of factor loadings of self-regulation items across samples

	Range of factor loadings by self-regulation skill							
	Goal-Setting[a]		Self-Esteem[b]		Emotional Control & Self-Monitoring[c]		Task Monitoring, Planning, & Initiation[c]	
Sample	Min	Max	Min	Max	Min	Max	Min	Max
All	**0.42**	**0.84**	**0.50**	**0.85**	**0.69**	**0.81**	**0.58**	**0.79**
Program								
FaDSS	0.47	0.85	0.52	0.84	0.70	0.80	0.57	0.78
Goal4 It!	0.46	0.83	0.50	0.86	0.69	0.87	0.64	0.87
LIFT	0.32	0.80	0.44	0.86	0.65	0.81	0.55	0.78
Gender								
Female	0.42	0.85	0.50	0.86	0.69	0.81	0.58	0.79
Male[d]	0.40	1.14	0.39	0.75	0.61	0.85	0.58	0.82
Age								
Less than or equal to 30 years old	0.45	0.83	0.50	0.87	0.70	0.82	0.58	0.77
More than 30 years old	0.38	0.87	0.50	0.85	0.68	0.81	0.59	0.82
Race and ethnicity								
Hispanic	0.55	0.81	0.46	0.84	0.64	0.80	0.61	0.83
Black	0.24	0.84	0.46	0.88	0.62	0.80	0.52	0.79
White, other	0.41	0.86	0.54	0.85	0.70	0.84	0.60	0.80
Education								
Less than high school diploma or GED	0.37	0.89	0.46	0.87	0.63	0.81	0.62	0.81
High school diploma or GED	0.50	0.85	0.48	0.87	0.69	0.81	0.57	0.79
College	0.39	0.88	0.54	0.86	0.70	0.84	0.56	0.82
Employment								
Worked for pay in past 30 days	0.55	0.89	0.54	0.85	0.68	0.81	0.60	0.81
Did not work for pay in past 30 days	0.25	0.93	0.45	0.88	0.69	0.82	0.54	0.77

Source: Evaluation of Employment Coaching baseline survey.

[a] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Goal-Setting skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[b] A 0- to 3-point scale based on the extent to which respondents agree with statements that reflect a high level of Self-Esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

[c] A 0- to 2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

[d] Although the model estimation converged successfully, the estimate of one of the factor loadings exceeded one for the Goal-Setting measure for male sample members. For that reason, we caution interpretation of this model.

Table A.4. Overall fit statistics across samples

	Fit statistics				
	Root mean square error of approximation (RMSEA)				
Sample	Estimate	95 percent confidence interval (lower and upper bounds)		Comparative Fit Index (CFI)	Tucker Lewis Index (TLI)
All	**0.06**	**0.06**	**0.06**	**0.94**	**0.93**
Program					
FaDSS	0.06	0.06	0.06	0.94	0.93
Goal4 It!	0.06	0.06	0.07	0.95	0.95
LIFT	0.05	0.05	0.05	0.95	0.94
Gender					
Female	0.06	0.06	0.06	0.94	0.93
Male[d]	0.05	0.04	0.06	0.94	0.94
Age					
Less than or equal to 30 years old	0.06	0.06	0.06	0.93	0.93
More than 30 years old	0.05	0.05	0.06	0.95	0.95
Race and ethnicity					
Hispanic	0.05	0.05	0.06	0.95	0.95
Black	0.05	0.04	0.05	0.95	0.94
White, other	0.06	0.06	0.07	0.94	0.93
Education					
Less than high school diploma or GED	0.05	0.05	0.06	0.95	0.94
High school diploma or GED	0.06	0.06	0.07	0.93	0.92
College	0.05	0.05	0.06	0.95	0.95
Employment					
Worked for pay in past 30 days	0.06	0.05	0.06	0.95	0.94
Did not work for pay in past 30 days	0.06	0.06	0.06	0.94	0.93

Source: Evaluation of Employment Coaching baseline survey.

[a] Although the model estimation converged successfully, the estimate of one of the factor loadings exceeded one for the Goal-Setting measure. For that reason, we caution interpretation of this model.

Appendix B: Formula for power calculations

We extended the standard formula for calculating minimum detectable effects (MDE) to incorporate measurement error. As described in Schochet (2008), the MDE for randomized controlled trials with individual-level random assignment is typically calculated as:

$$MDE = \left[T^{-1}\left(1-\frac{a}{2}\right)+T^{-1}(b)\right]\times\frac{SE}{\sigma_\theta},$$

where T^{-1} is the inverse of the student's *t* distribution with *df* degrees of freedom ($N - 2$), $\boldsymbol{a}$ is the significance level, *b* is the power, *SE* is the standard error of the estimated impact, and σ_θ is the standard deviation of the outcome (in this case, the true self-regulation skill measured without error).

With measurement error, the standard error may be written as:

$$SE = \sqrt{\frac{\sigma_M^2}{Np(1-p)}},$$

where *N* is the sample size, *p* is the probability of random assignment, and $\boldsymbol{\sigma_M^2}$ is the variance of the measure of the self-regulation skill (including measurement error) that is calculated as the average across *K* items.

We assume that the items comprising the measure take the form:

$$M_k = \theta + \varepsilon_k,$$

where θ is the true skill and ε_k is the measurement error for item *k*. Assuming that the measurement error for each item has the same variance (σ_ε^2), that the measurement error for a given item (ε_k) is independent of the true skill (θ) and that the measurement error is independent across items, it can be shown that the inter-item correlation *r* may be written as:

$$r = \frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\varepsilon^2},$$

which implies that $\sigma_\varepsilon^2 = \frac{(1-r)\sigma_\theta^2}{r}$. Using this fact, the variance of the measure defined as the average across items ($M = \sum_{k=1}^{K} M_k/K$) may be written as:

$$\sigma_M^2 = \sigma_\theta^2 + \frac{\sigma_\varepsilon^2}{K} = \sigma_\theta^2\frac{r(K-1)+1}{Kr}.$$

Combining this expression with the standard MDE formula yields the following:

$$MDE = \left[T^{-1}\left(1-\frac{a}{2}\right)+T^{-1}(b)\right]\times\sqrt{\frac{r(K-1)+1}{rKNp(1-p)}}.$$

Note that the elements of the formula may be approximated during the planning of a study if estimates of Cronbach's alpha are available for a given measure. With measures that have the same standard deviation and the same inter-item correlation for all pairs of items, Cronbach's alpha may be written as:

$$\alpha = \frac{Kr}{1+(K-1)r},$$

where *r* is the average inter-item correlation (Nunnally and Bernstein 1994). Therefore, it is possible to approximate the inter-item correlation *r* from estimates of α and the original number of items in a measure. Using the approximation of *r*, an analyst could then use the formula for the MDE to estimate statistical power for different numbers of items in the measure.

In addition, the MDE may be written in terms of α, as follows:

$$MDE = \left[T^{-1}\left(1-\frac{a}{2}\right)+T^{-1}(b)\right]\times\sqrt{\frac{1}{\alpha Np(1-p)}}.$$

OPRE

Mathematica
Progress Together