SUBSIDIZED AND TRANSITIONAL EMPLOYMENT DEMONSTRATION (STED) AND ENHANCED TRANSITIONAL JOBS DEMONSTRATION (ETJD)

SUPPORTING STATEMENT B
REQUEST FOR OMB CLEARANCE

OMB No.: 0970-0413

Submitted By:

Office of Planning, Research and Evaluation

Administration for Children and Families

U.S. Department of Health and Human Services

7th Floor, West Aerospace Building

370 L'Enfant Promenade, SW

Washington, D.C. 20447

October 2012

Revised: May 2013

Revised to include alternate surveys instruments, which were developed for the youth/young adult STED sites. All changes from the original approved package are highlighted.

B. COLLECTION OF INFORMATION USING STATISTICAL METHODS

B1. Respondent Universe, Sampling Selection, and Expected Response Rates

This section focuses on our sampling plans for the follow-up surveys. Plans for interviews and questionnaires for the implementation analysis (which will not be analyzed for statistical differences) are discussed in Part A.

Each project plans to include a total of seven evaluation sites. However, because two of the ETJD sites will be evaluated under STED, there will be a total of 12 sites in the two projects combined. ACF and ETA estimate that 1,000 individuals will be randomly assigned at each site, for a total of 12,000 in the study across the two projects. In each site, 500 of these individuals will be assigned to the treatment group and 500 will be assigned to the control group.

As Exhibit 2.1 shows, both the 6-month and 12-month surveys will be administered to all sample group members in the STED evaluation; only the 12-month survey will be administered to the sample members in the ETJD sites. As discussed later, extensive efforts will be taken to contact all sample group members as the target response rate for both surveys is 80% of the research sample at each site. Thus, for the 6-month survey, the total sample size across all seven STED sites (including the two ETJD sites that are also in the STED evaluation) is 7,000 with an expected number of respondents equal to 5,600. For the 12-month survey, the total sample size across all sites is 12,000, with 9,600 expected respondents.

Exhibit 2.1
Follow-up Survey Sample Sizes

Survey Efforts/Sites	Sites	Sample size per site		
		Research Sample	Survey Sample	Survey Respondents
6-Month Survey	STED sites (7)	1,000	1,000	800
12-Month Survey	All sites (12)	1,000	1,000	800

A fuller accounting of sample sizes, along with a complete list of data collection instruments in this submission, is outlined in Exhibit 1.2, Annual Burden Estimates, in Part A of the Supporting Statement.

B2. Procedures for Data Collection and Statistical Analysis

The 6- and 12-month follow-up survey data will be collected through a mixture of telephone and in-person outreach and interviewing strategies to maximize response rates. The timing of the data collection efforts was determined by the research questions motivating each survey effort. That is, the 6-month surveys are focused on the immediate, non-financial benefits of employment and thus the timing of survey administration is designed to collect information while or shortly afterwards participation in the STED programs. Likewise, the 12-month surveys are designed to measure post-program outcomes and, therefore, the timing of the survey administration is designed to collect information shortly after program participation has concluded.

Both the 6- and 12-month survey data will be used to estimate program impacts. The basic procedure for estimation of program impacts will be to compare the average outcomes of program and control group members. These estimates will be calculated using multivariate regression models that predict outcomes as a function of assignment to the program group and participant baseline characteristics. Controlling for baseline characteristics will increase the statistical precision of the impact estimates for a given sample size, neutralize chance differences in characteristics between the program and control groups, and reduce attrition bias from missing data.

A strength of random assignment is that it is easy for nontechnical audiences to understand. The evaluation team will therefore emphasize methods that are appropriate and straightforward. The primary analytical method will be comparisons of average outcomes for program group members (regardless of attrition from program participation) and control group members, and comparisons of distributions of outcomes for program and control group members.

The general form of the regression models which will be used to estimate program impacts is as follows:

$$Yi = \alpha + \beta Pi + \delta Xi + \varepsilon i$$

where

Y_i is the outcome measure for sample member i;

P_i equals one for program group members and zero for control group members;

X_i is a set of background characteristics for sample member i; and

ε_i is a random error term for sample member i.

The coefficient β is interpreted as the impact of the program on the outcome. The regression coefficients, δ, reflect the influence of background characteristics. The functional form and estimation method will depend on the scale of measurement of the outcome for which impacts are estimates; for example, continuous outcomes will be estimated using ordinary least squares (OLS) regression.

Standard statistical tests such as the two-group t-test (for continuous variables such as earnings) or chi-square tests (for categorical measures, such as educational attainment) will be used to determine whether estimated effects are statistically significant, after adjusting for differences in characteristics between the program and comparison groups at the 1%, 5%, or 10% level. We expect to use regression adjustment to increase the power of statistical tests that are performed, although we will perform checks to ensure that regression adjustment does not significantly change the estimated impacts of the interventions. In order to reduce multiple test bias, outcomes will be pre-specified as primary versus secondary and we will strive to keep the number of comparisons as small as possible.

Subgroup analysis. Impacts will be calculated for key subgroups to better understand what works best for whom. In MDRC studies, subgroup impacts have been estimated several different ways. In "split-sample" subgroup analyses, the full sample is divided into two or more mutually exclusive and exhaustive groups (for example, by gender or for those with more versus less work experience at the time of random assignment). In this approach, impacts are estimated for each group separately. In addition to determining whether the intervention had statistically significant effects for each subgroup, tests will be conducted to determine whether impacts differ significantly across subgroups. For STED and ETJD we will be particularly interested in how results vary previous labor market experience and level of disadvantage.

We will strive to keep subgroup comparisons to a minimum number for which theory and prior studies provide good reasons for expecting subgroup differences on employment outcomes. This is to guard against the chance of a "false positive", which stems from the fact that the more subgroups that are examined, the greater the chance of finding one with a large effect, even when there are no real differences in impacts across subgroups.

Exhibit 2.2 reports the estimated minimum detectable effects (MDEs) for the 6- and 12-month survey given the planned sample sizes and response rate. In this case, the MDE is the smallest true effect that would generate statistically significant impacts in 80 percent of evaluations with a given sample size. Because the ETJD and STED programs and populations might differ substantially from site to site, it is important that we have the capability to detect reasonably sized impacts in each of the sites. Also note that, as the sample size column indicates, the respondent sample for the survey will be 800 per site (based on assumption of an 80 percent response rate among a fielded sample of 1,000).

Exhibit 2.2

Minimum Detectable Effects, Per Site, Per Survey

	Respondent Sample	Fielded Sample
Total sample size (2 group sites)	800	1,000
Sample size per research group	400	500
Minimum Detectable Effects		
Arrested, Year 1	8.1	7.3
Convicted, Year 1	6.1	5.4
Incarcerated, Year 1	7.4	6.6
Employed at interview	8.4	7.5
Total earnings, Year 1	961	859
Self-reported drug use or tested drug use	8.5	7.6
Average Welfare Receipt Payments, Six quarters	505	452
Ever Paid Child Support, end Year 1	7.5	6.7
Maximum MDE with sample size (Std. Dev. = 0.5)	8.5	7.6

NOTE: MDEs are for two-tailed tests at 0.1 significance with 80 percent power using fixed effects site estimates and no covariates. The following assumptions were made regarding control group proportions (based on related projects): Arrests: 35 percent; Convictions: 15 percent; Incarcerations: 25 percent; Employment: 59 percent; Drug use: 48 percent; Child Support Payments: 26 percent. For earnings, we assumed a standard deviation of $5,000. For average welfare receipt payments, we assumed a standard deviation of $2,962.

As the table shows, for the proposed site survey sample, MDEs for percentage outcomes measured with the survey range from about 6 to 8.5 percentage points, depending on the outcome. For the full administrative records samples, MDEs would range from approximately 5 to 7.5 percentage points. The table also shows impacts on earnings and welfare payments. For this example, we assumed a control group Year 1 earnings level of approximately $5,000 (this was based on some of our recent ex-offenders studies). MDEs for earnings range from $961 (in the survey sample) down to $859 (in the full research sample). Thus, the planned sample size and anticipated response rate will allow us to detect policy-relevant impacts at the site level.[1]

Several of the survey items were adopted from existing scales. Where scales are used, we will assess the reliability of the scale for the STED/ETJD samples. The source for the general self efficacy scale, used in the six-month survey is Schwarzer & Jerusalem (1995).[2] This scale has been used internationally for several years and a sampling across 23 nations found that Cronbach's alphas ranged from .76 to .90, with the majority in the high .8 range. Regarding

[1] The evaluation will also include a cost-benefit analysis. To estimate the program costs, the evaluation team will collect financial reports from each site. They will select a period approximately one year after the program began operations. Additionally, a staff time study will be administered to all program staff and will be used to allocate program costs across key program components. The cost-benefit analysis draws on the cost analysis and the analysis of program impacts.

[2] "Generalized Self-Efficacy scale," Schwarzer, R., & Jerusalem, M. (1995). In J. Weinman, S. Wright, & M. Johnston, *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35-37). For more information see: http://userpage.fu-berlin.de/~health/faq_gse.pdf.

validity, the authors report that "Criterion-related validity is documented in numerous correlation studies where positive coefficients were found with favorable emotions, dispositional optimism, and work satisfaction"[3,4].

Other scales used in the surveys:

- The emotional support scale is from RAND.[5] Quoting this source: "Multitrait scaling analyses supported the dimensionality of four functional support scales (emotional/informational, tangible, affectionate, and positive social interaction) and the construction of an overall functional social support index. These support measures are distinct from structural measures of social support and from related health measures. They are reliable (all Alphas >0.91), and are fairly stable over time. Selected construct validity hypotheses were supported."

- Questions MPH1-MPH2 are from the RAND "36-Item Health Survey 1.0 Questionnaire."[6] Reliability (measured via Cronbach's alpha) on the subscales ranges from .78 up to .93.

- Domain Specific control is from the Health and Retirement survey from the University of Michigan.[7]

- Regarding the material support scales, question MSS1 is sourced from the "The Making Connections Cross-Site Survey," Annie E. Casey Foundation. Questions MSS2-MSS3 were sourced from "The Wisconsin Longitudinal Survey," The Center for Demography of Health and Aging (CDHA) at the University of Wisconsin-Madison

- The Social Network Roster and Relationship Origin (questions J1 and J2) are from "Personal Networks and. Community Survey," Princeton Survey Research Associates International.

---

[3] http://userpage.fu-berlin.de/health/engscal.htm

[4] Updated validity information is shown in: Updated psychometric findings have been published recently, for example, in: Scholz, U., Gutiérrez-Doña, B., Sud, S., & Schwarzer, R. (2002). Is general self-efficacy a universal construct? Psychometric findings from 25 countries. *European Journal of Psychological Assessment, 18*(3), 242-251.

[5] http://www.rand.org/pubs/reprints/RP218.html and http://cmcd.sph.umich.edu/assets/files/Repository/Women%20Take%20Pride/The%20MOS%20Social%20Support%20Survey.pdf.

[6] For more information, please see http://www.rand.org/content/dam/rand/www/external/health/surveys_tools/mos/mos_core_36item_scoring.pdf

[7] Clarke, Philippa, Gwenith G. Fisher, Jim House, Jacqui Smith, and David R. Weir. Guide to Content of the HRS Psychosocial Leave-Behind Participant Lifestyle Questionnaires: 2004 & 2006 (2008).

- D3 is the K6 scale (Kessler, et al., 2003) and is designed to discriminate case of serious mental illness from non-cases. It was developed for use in the U.S. National Health Interview Survey with support from the National Center for Health Statistics.

- The Rosenberg Self-Esteem scale (Young Adult Instruments only; 6-month item B3, 12-month item E5) is a widely used instrument developed by Morris Rosenberg[8] in the mid-sixties and has been extensively tested and validated for reliability[9].

- The Career Commitment Measure (Young Adult Instruments only; 6-month item C2, 12-month item E7) was developed by Carson and Bedeian[10] and has been used successfully with a similar population[11]. The scale has been assigned for reliabilities (alphas range from .79 to .85), discriminant validity, and construct validity.

B3. Maximizing Response Rates and Issues of Nonresponse

The goal will be to achieve an 80 percent response rate for both surveys at each site (STED and ETJD) included in the survey effort. Procedures for obtaining the maximum degree of cooperation and thus the response rate include:

- Maximize use of contact information collected by the program at the point of random assignment, including email addresses and alternate contact information for at least three other individuals whom the respondent identified as likely to know how to find him or her;

- Using advance letters, greeting cards, and email contacts (See Appendix E);

- Conveying the purposes of the survey to respondents so they will thoroughly understand the purposes of the survey and perceive that cooperating is worthwhile;

- Providing a toll-free number for respondents to use to update their contact

[8] Rosenberg, Morris. (1965) Society and the Adolescent Self-Image. Princeton, NY: Princeton University Press.

[9] Blascovich, Jim and Tomaka, Joseph. 1993. "Measures of Self-Esteem." in Robinson, J.P., et al (editors), Measures of Personality and Social Psychological Attitudes, Third Edition. Ann Arbor: Institute for Social Research.

[10] Carson, Kerry D. and Bedeian, Arthur G. 1994. "Career Commitment: Construction of a Measure and Examination of its Psychometric Properties." Journal of Vocational Behavior, 44, 237-262.

[11] Personal Responsibility Education Program (PREP) Multi-Component Evaluation Survey (OMB Control No. 0970-0398)

information in anticipation of the survey;

- Training site staff to be encouraging and supportive, and to provide assistance to participants as needed;
- Hiring interviewers who have necessary skills for encouraging cooperation;
- Implementing a tracking strategy that keeps in touch with the sample members and periodically requests updated contact information (see Appendix E).
- Training interviewers and field locators thoroughly in conversion and avoidance of refusals;
- Timing cases from the CATI center to tracking and the field so that each case will not remain in the CATI center for more than 30 days.
- Offering appropriate payments to participants for participating in the survey effort.

The follow-up surveys are designed to be administered in the home or by telephone. Once contacted, the interviewer will administer the survey over the telephone using the CATI questionnaire, or in-person using the CAPI questionnaire if attempts to reach the respondent via phone are not successful. This process is discussed more below.

Interviewers will also be trained to distinguish "soft" refusals from "hard" ones. Soft refusals often occur when the sample member has been reached at an inopportune time. In these cases, it is important to back off gracefully and to establish a convenient time to call or come back rather than to persist at the moment. Hard refusals do occur and must also be accepted gracefully by the interviewer.

Procedures for contacting hard to reach respondents

The survey firms – DIR and Abt/SRBI – telephone interviewers will first try to reach the sample member and administer the first follow-up survey using CATI. The telephone interviewers will be using the original contact information collected at baseline and provided to the survey firms by MDRC. An initial attempt will be made to reach the sample member, scheduling an appointment for completion through the CATI system if it is best for the respondent. If the number is no longer valid (out of service or reassigned to another person), then the interviewer will attempt to locate a new telephone number by calling directory assistance. If no new telephone number can be located for the respondent then the survey firms will try to update the number using a service offered by Lexis Nexis. Any new numbers will be loaded into the CATI system to be dialed by interviewers. The telephone interviewer may also call the numbers given for sample member's secondary contacts. These contacts were given to us by the sample member at baseline, as relatives or friends who do not live in the same household as the sample member but will always know how to reach them. Every attempt (call disposition) to contact the sample member or their secondary contacts and its outcome is recorded in CATI.

This information will be provided to the field interviewer once sample is transferred to the field in the form of a respondent contact sheet. The respondent contact sheet will be what the field interviewer uses to record and code all of their attempts to contact the respondent.

Once the telephone interviewers have exhausted all leads, the case will be transferred to the survey firms field interviewers to locate the sample member and administer the surveys using Computer Assisted Personal Interviewing (CAPI). The field interviewer will review all the notes and attempts from CATI in the respondent contact sheet. They will first try calling the respondent using any numbers believed to be working by the telephone interviewers. This is done because sometimes sample members do not answer calls from out of area but will answer a call from a local number. If none of the telephone numbers are useful, they will attempt to contact the sample member or their secondary contacts in person. If necessary, they may speak to neighbors of the sample member or their secondary contacts, or to others in the community, to find out if anyone knows the sample member's whereabouts. If all attempts to contact fail, we will conduct an advanced Lexis Nexis search which provides address, name and telephone history of the respondent. These searches are performed by the field managers. Field managers sift through this data and provide additional contact information to the interviewers. Based on prior experience with similar populations, it is anticipated that 57 percent of the completes will be obtained by telephone and the remaining 43 percent of the completes will be obtained in-person.

Viability of attaining the goal response rate

The survey firms – DIR and Abt/SRBI – have extensive experience managing multisite longitudinal field studies and attaining high response rates. These organizations employ professionally trained telephone interviewers experienced in obtaining high response rates and a nationwide roster of experienced field staff across the United States that are available to work on studies as they develop. Numerous MDRC studies with similar populations have achieved 80 percent response rates. For example, DIR recently achieved an 81 percent response rate for a sample which included ex-offenders (this was a 12-month follow-up survey for the Work Advancement and Support Center demonstration (Miller et al., 2012)). The Parents' Fair Share study, which included non-custodial parents, achieved a response rate of 78 percent (Miller & Knox, 2001). The Philadelphia Hard-to-Employ study (a transitional jobs program for TANF recipients) achieved a 79 percent response rate (Jacobs & Bloom, 2011). Several sites in the Employment Retention and Advancement evaluation achieved 80 percent response rates as well (Hendra et al., 2010).

Abt Associates and its survey subsidiary, Abt SRBI, have achieved among the highest survey response rates in the industry using a variety of methods specifically aimed at maximizing responses for large-scale studies with difficult-to-track populations. Abt's work on the Supporting Healthy Marriage project (for MDRC), the Survey of Recently Naturalized Citizens for U.S. Citizenship and Immigration Services, and the Veterans Employability Research Study for the Department of Veterans Affairs involves multi-site, large-scale, mixed-mode surveys that require extensive tracking efforts.

We will monitor survey completion rates within the sample cohort (defined by time of random assignment) by research group, site, and sub-population to provide feedback to the survey firms regarding the need to focus or intensify recruitment efforts.

Assessing and correcting for survey nonresponse bias.

Survey nonresponse can bias the impact estimates if the outcomes of survey respondents and nonrespondents differ, or if the types of individuals who respond to the surveys differ across the program and control groups. The safest and best way to avoid or reduce this problem is, of course, to maximize response rates to the survey, and we have proposed methods that we believe will do so. Despite these efforts, however, it is certain that we will not achieve a 100 percent response rate and, in fact, that a reasonable proportion of sample members will not complete the survey, leading to the potential for nonresponse bias to affect the survey results and, thus, the impact estimates. We will use several methods to assess the effects of survey nonresponse during data collection and using data collected for the study.

During data collection, we will take steps to understand, monitor, manage and address potential sources of non-response bias. During the survey fielding period, we will receive weekly reports from our survey contractors providing information on contact attempts and disposition status which will enable us to monitor response rates by sample cohort (defined by time of random assignment), research group, site, and target population (i.e., Non-Custodial Parents, Ex-Offenders, TANF Recipients, etc.). We will also monitor response for specific sub-populations of the sample who may have barriers to participation in the survey effort, including (but not limited to) non-English speakers and incarcerated sample members. Should significant gaps in response rates among these groups occur, we will intensify recruitment efforts for the affected group. These intensified efforts will include prioritizing the efforts of the most experienced survey interviewers towards the affected group and increasing the use of local interviewers to locate and recruit participants.

We will also examine nonresponse using data collected for the study. First, we will use baseline data (which will be available for the *full* research sample) to conduct statistical tests (chi-squared and t-tests) to gauge whether treatments who respond to the interviews are fully representative of all treatment group members, and similarly for control group members. Noticeable differences in the characteristics of survey respondents and nonrespondents could suggest the presence of nonresponse bias. Furthermore, we will test whether the baseline characteristics of *respondents* in the two research groups differ from each other. Although baseline characteristics for the full sample should not differ much between the program and control groups, significant differences between program and control group respondents could mean that impacts estimated from surveys will confound program impacts with pre-existing differences between the groups.

Second, we will assess nonresponse bias using administrative records data. For example, we will examine whether *impacts* on arrests or employment rates differ for survey respondents and survey nonrespondents. If program impacts are substantially different for respondents and nonrespondents, that would make us more cautious about drawing conclusions from the survey.

We will use several approaches to correct for potential nonresponse bias in the estimation of program impacts. First, as discussed, we will adjust for observed differences between program

and control group respondents using regression models. Second, because this regression procedure will not correct for differences between respondents and nonrespondents in each research group, we will construct sample weights so that the weighted observable baseline characteristics of respondents are similar to the baseline characteristics of the full sample of respondents and nonrespondents. We will construct weights for program and control group members using the following three steps:

1. Estimate a logit model predicting interview response. The binary variable indicating whether or not a sample member is a respondent to the instrument will be regressed on baseline measures.
2. Calculate a propensity score for each individual in the full sample. This score is the predicted probability that a sample member is a respondent, and will be constructed using the parameter estimates from the logit regression model and the person's baseline characteristics. Individuals with large propensity scores are likely to be respondents, whereas those with small propensity scores are likely to be nonrespondents.
3. Construct nonresponse weights using the propensity scores. Individuals will be ranked by the size of their propensity scores, and divided into several groups of equal size. The weight for a sample member will be inversely proportional to the mean propensity score of the group to which the person is assigned.

This propensity score procedure will yield large weights for those with characteristics that are associated with low response rates (that is, for those with small propensity scores). Similarly, the procedure will yield small weights for those with characteristics that are associated with high response rates. Thus, the weighted characteristics of respondents should be similar, on average, to the characteristics of the entire research sample.

It is important to note that the use of weights and regression models adjusts only for *observable* differences between survey respondents and nonrespondents in the two research groups. The procedure does not adjust for potential unobservable differences between the groups. Thus, our procedures will only partially adjust for potential nonresponse bias. We will use administrative data to assess whether such bias is present in our data, as discussed above.

B4. Pre-Testing

Many of the questions proposed for this survey are either identical to questions used in prior evaluations or are similar, if not identical, to questions used in previous national surveys. Consequently, many of the items and measures have been thoroughly tested on larger samples.

MDRC will work closely with DIR, Inc. and Abt SRBI's senior staff to conduct formal pretests of both the 6-month and 12-month follow-up surveys, with a convenience sample that are not included in the survey sample. Because the sample for the pilot test will include only nine or fewer study participants, our understanding is that this effort does not require a separate OMB review and approval process, and these hours are not included in our burden estimates. These

pretests will provide more definitive estimates about the length of the surveys and their various components, as well as lead to improvements in questions, introduction scripts, wording and document formatting. Following the pretests, respondents will be debriefed about the clarity of the questions and any potential problems with the instruments. Interviewers will also be debriefed concerning any problems they encountered in the survey – and they will recommend improvements. The survey instrument will be revised to incorporate the survey firms' recommendations for improving the readability of questions that respondents had difficulty understanding. If revisions occur, updated instruments will be submitted to OMB. However, given that most of the questions are from existing surveys, we do not expect many changes in the instruments after piloting. Each survey will be translated into Spanish versions once the English versions are finalized.

B5. Individuals Consulted on Statistical Aspects and Individuals Collecting and/or Analyzing Data

The information for the STED and ETJD studies is being collected by MDRC and its subcontractors, Branch Associates, DIR, MEF Associates, and Abt Associates on behalf of ACF and DOL. With ACF and DOL oversight, MDRC and its subcontractors were responsible for developing the instruments.

ACF/OPRE Contact:
Girley Wright
(202) 401-5070
Girley.wright@acf.hhs.gov

DOL/ETA Contact:
Eileen Pederson
(202) 693-3647