ED Response to OMB Comments Regarding Evaluation of the Teacher Incentive Fund (TIF) Program: Data Collection Instruments, 1875-New

The order of questions has been rearranged because we believe that addressing question five first will make the other answers easier to understand.

5. We would like to understand the "big picture" for this study - what are the possible phases and how do they fit together?

This evaluation has three closely related parts.

The first part of the evaluation (research questions 1-4) is an implementation study. To learn about TIF implementation we propose to use:

- Phone interviews of a small sample of respondents from each grantee. These will be used to provide up-to-date objective descriptions of the grantees and their program implementation. Because of the small sample size from each respondent, these will be insufficient for making generalizable claims on aspects of programs.
- Two rounds of case studies from a larger sample of respondents from 12 grantees. These will provide a more nuanced understanding of a subset of grantees. (Sampling for these is described in the original OMB package.) These will help the research team to establish hypotheses about the potential relationships between program design, implementation, grantee context, and educator motivation. These hypotheses will, in turn, be subject to more objective and generalizable examination through surveys and an optional outcomes analysis.
- A teacher survey (to clarify, this would be a paper survey not a phone survey) of a random sample of teachers from each grantee. (This survey has been added to evaluation by a recent modification to the contract and will be submitted for OMB approval). The teacher survey will provide generalizable, descriptive data on the programs, implementation, and teachers' reports on their subsequent motivation (which cannot be gleaned from any extant or objective measure).
- A principal survey of a random sample of principals from each grantee. (This survey was also added to the evaluation by a recent modification to the contract. The survey will be submitted for OMB approval soon). The principal survey will provide generalizable, descriptive data on the programs, implementation, and principals' reports on their subsequent motivation (which cannot be gleaned from any extant or objective measure).

The second part of the evaluation is a feasibility study (research question 5). The purpose of this study is to examine a range of possible designs for an outcomes analysis. For each proposed design we will investigate the possibility of gathering the necessary data, and the strengths and weaknesses of various design options given the data that are available. As part of the feasibility study, we will propose various ways to address research questions 6-10. The feasibility study will include an investigation of possible comparison groups, and examination of the quality of available measures, appropriate power analyses, and a discussion of the internal and external validity of various designs. The three design types being considered are a regression discontinuity design, a difference in differences design, and a meta-analysis.

The third part of the evaluation is an optional outcomes analysis. The majority of the questions, 6-9, require using extant data to explore the relationship between grantee's participation in TIF and a range of educator and student outcome measures. For question 10, which asks about the relationship between performance pay design models and outcomes, information from the

implementation study on program characteristics will be used to generate grantee-level data to include in analyses. The research team is aware that various designs lend themselves more (randomized trial) or less (less rigorous quasi-experiments) to causal inference. We will not make inferences that are inappropriate to the selected research design(s).

1. **We request that ED come back to OMB for additional review before exercising the contract option for the impact study with the results of the feasibility and the contractor's proposed evaluation design. Please let us know if this is not acceptable.**

Yes, we will submit the results of the feasibility study to OMB for additional review. Changes have been made to the supporting statements to make our assumption explicit.

2. **In looking at how successful TIF grantees are in attracting and retaining effective principals and teachers (research question #6), how will the comparison group ("similar schools") be defined?**

This is one of the key questions in the feasibility study. Different possible designs have different comparison groups. Prior to conducting an outcomes analysis, we will (as indicated above) submit details on the comparison groups to OMB.

3. **How will the study determine whether the performance evaluations available are of sufficient quality (valid and reliable) to be used to draw conclusions about the improved performance of teachers (research question #8)?**

The answer to this question is also part of the feasibility study and will be discussed in a later OMB submission before the optional outcomes analysis is exercised. As part of the phone interviews, we will gather basic descriptive data on the evaluation systems that will be used in the feasibility study to begin to learn about the evaluation systems.

4. **What methodology will the study use for answering research question #9?**

Addressing this question (in addition to investigating how to answer questions 6, 7, 8, and 10) is the purpose of the feasibility study. At this point, the feasibility study is investigating the possibility of using one or more of the following design types: regression discontinuity design, difference in differences (including interrupted time series), and meta-analysis of local evaluations.

6. **Please change the "voice" of the supporting statement to that of the Department, rather than the contractor (page 9, 1st and 3rd paragraphs; page 5, part b last paragraph, etc.).**

In writing the OMB package the intention was to reflect the Department's evaluation, however you have pointed out an inconsistency in our writing. There are multiple occasions when ED is referred to explicitly as a separate entity and the "we" in our original submission therefore only reflected the research team of SRI, Urban, and BPA. In this document, we have clarified that "we" means ED and the "researchers" or "research team" refers to SRI, UI and BPA. We believe that this description of ED's specific role at certain decision points and processes created some confusion (see response to question 13 below). We have read through the supporting statement and in all cases now write exclusively from the "voice" of ED in the revised Supporting Statements A and B and in our response to your questions.

7. **Please provide more information about the analysis plans for the collected data, especially on that collected to answer evaluation questions 6 through 9 (on page 3, SS A). We are particularly concerned that any "outcome analysis" not be used to imply causality prior to the conduct of a more rigorous evaluation.**

This is part of the outcomes tasks, which are currently optional tasks. Before these tasks are exercised, we will provide appropriate information to OMB. However, as stated previously, we are aware of the limitations of non-experimental designs for causal inference.

8. **What are the specific criteria that ED will use to determine feasibility for a rigorous evaluation?**

The research team will present to ED information on data availability (including likely quality of measures), power analyses, possible research designs (and the threats to their validity), and costs. ED will weigh the costs against the likely inferences supported by the designs (which can be determined based on available data, power, and precise design) in deciding which design(s) are feasible and desirable.

9. **Confidentiality - please clarify under which statute confidentiality assurances are being made. If not, please do not use the phrase "we will assure their confidentiality to the extent offered by law" in SS a 10 or in other materials. Rather, rely on an explanation of procedures designs to minimize the risk of inadvertent disclosure etc.**

As the lead in the data collection of the evaluation, the research team will adhere to the Multiple Projects Assurance with the Office of Protection from Research Risks (OPRR) maintained by SRI. SRI's Assurance number is M-1088. SRI's Human Subjects Committee is its official Institutional Review Board (IRB) charged with responsibility for the review and approval of all research involving human subjects. SRI clears all data collection protocols through its internal Human Subjects Committee as a safeguard to protect the rights of our research subjects. The National Evaluation of the Teacher Incentive Fund has already gained approval from SRI's Human Subjects Committee.

The risks and benefits of participating in this study, which are expected to be nominal, will be explained to potential participants before we begin the interviews. Informants' consent will be actively requested and documented with a consent form. We will protect our participants' confidentiality. Only research staff on this project will have access to actual interview responses, and those data, including the audiotape of the interview, will be stored on secure computers. The audiotape will be destroyed at the conclusion of the study.

Please see the revised explanation in SS A #10.

10. **Voluntary v. mandatory - SS B3 indicates that grantees were given extra points for indicating that they would participate in this evaluation. Did any awardees decline? If so, what are the implications? Does that mean this activity is mandatory for some grantees and voluntary for others? Is it mandatory of voluntary for those other than the Project staff? Please clarify.**

All grantees indicated a willingness to participate in the federal evaluation in their proposal. We interpret SRI's IRB clearance to indicate that participation in the phone interviews and case studies is optional for every individual, including project staff. We have attached IC1 and IC15.

11. Please clarify the specific activities that will occur during the 3 days of data collection per site in the case studies. Is all of the data collection via "interviews" or is some also via "observation?" If the latter, what and who will be observed?

All data will be collected through interviews.

12. Sampling:

An overarching challenge in describing the samples for the qualitative data collection (phone interviews and two rounds of case studies) is the variation in grantees. For example, one grantee is a single charter school, others are states, districts, consortia or charter schools. Additionally, in some grantees all schools within a district or districts within a state participate; in others a subset of schools or districts within the grantee unit participate. As a result, we cannot uniformly describe the governance structure across grantees or the participation of stakeholders in the process of developing and implementing the grants. Below we describe the uniform principles that will be used to identify respondents and share examples of how uniform principles may lead to differing sample sizes and precise respondents due to the variations in the nature of grantees.

a. Understanding that there is variation by site, please clarify the approximate universe size of each "informant" group at each site (e.g., TIF project staff). In general, is SRI planning to interview "the universe" or a sample? How was the sample size derived?

We plan to interview a sample of informants, which will be determined by the size and complexity of the grantee's structure. For example, in the case of the single charter school, the sample size for phone interviews might total about six respondents; in contrast for the state grantee with four large urban districts participating, a much larger sample size might be necessary (perhaps 16). The sample size will be driven partially by the specificity of roles that respondents may have. In a single school, the principal might also be in charge of all human resources tasks, however they might contract with an outside firm to manage the student achievement data and award calculations. Grantees proposals will be used to generate preliminary lists of desired respondents; project director interviews will be used to confirm the grantee's organizational structure, and thus to finalize the interview lists (i.e., sample size). Similar principles will govern sample size for case studies. Exhibit A1 of Supporting Statement A provides the total sample size for each information group; the average sample size can be calculated by dividing the total by 34, but there is substantial variation around the mean for reasons described above.

b. Is there a maximum number of interviewees in each category by site? Is there a minimum number by category? What is the rationale for both?

No. Due to the variation in grantees (see above) and the extent to which they involve various stakeholders in the process, there are no minimum or maximum that can be determined in advance. We anticipate the total sample size will likely range from 6 to 16, depending on the number of respondents who personally covered multiple respondent categories and the structure of the grantee (which may or may not necessitate multiple respondents within each category).

c. Please clarify how each project director will be directed to identify relevant participants from each group. For example, will SRI seek a "universe" list of

educators and then sample from it or will it ask for recommendations of several educators that meet certain criteria?

The research team SRI will develop initial lists of desired respondents based on the analysis of extant documents from each grantee. To select educators (principals and teachers who were not involved in the planning or rollout of the grants) the research team will analyze the variation within the grantee (e.g., is it composed of a few similar schools, or is there substantial variation in school characteristics within a grantee) and the design of the incentive program (with special attention to whether all educators participate and whether all appear equally likely to be able to achieve an award). In the case where there is variation, researchers might request that the project director help in the identification of several schools within researcher-determined strata and might then randomly select from eligible schools. A similar process could be used to select educators within schools.

d. Whatever the selection method, how concerns about bias be addressed?

The sampling procedures described in c should reduce the likely bias of educator respondents. However, project staff will form the majority of the interviews, especially for the phone survey (less for the case studies). We recognize that these staff members are likely to be supportive of the programs. To counter this bias, we have received a contract modification to include teacher and principal surveys as part of our implementation study (the subject of a later OMB package), which will have a large, randomly selected samples.

e. How does the teacher sample size (approximately 4 and 18 per site respectively for the telephone and site visit interviews) compare with the anticipated sample size of the forthcoming telephone survey? Given the telephone survey, why is the sample size so large for the first telephone interview and site visits?

There is no forthcoming telephone survey. There are forthcoming paper surveys (requiring approximately 10-15 minutes per respondent), with proposed sample size of several thousand educators (to be justified fully based on power analyses in the forthcoming OMB package). Given the complexity of grantee organizations, the phone interviews and case studies have small samples designed to provide multiple perspectives on some issues and triangulation of broad issues. A sample size of 4 teachers would only be necessary for phone interviews in places where teachers also assumed other roles (e.g., on the grantee planning committee, union representative). We do not anticipate selecting more than four teachers in any grantee within the “teacher” respondent category.

Similar principles will apply for the selection of case study respondents, where the other roles teachers play and the number of schools selected due to grantee size and within-grantee diversity will drive the number of teachers sampled.

f. Conversely, what is the rationale behind interviewing only about one "stakeholder" per TIF site? How will this list be generated and what is the rationale for the sample size?

The primary purpose of the phone interviews is to provide an up-to-date description of the program design and implementation. Stakeholders will be interviewed when they can be good informants about those topics and their perspectives are not duplicative of others already on the informant list. For example the definition of stakeholder also includes representatives of the media, which will be highly informed about the grantees (and therefore useful informants) in some places and less knowledgeable (and therefore, not included) in others. For yet other

grantees (e.g., small rural areas, single charter school, charter consortium) it is possible that there will be no real local media presence that is knowledgeable about the TIF grant. These examples highlight the variation in possible media that would lead to their presence or absence on our list of respondents for a given grantee.

Additionally, we account for the fact that some stakeholder representatives will be sampled as respondents in other categories (so a single interview will cover multiple respondent categories. We have not double-counted respondents in our estimated sample size). For example, in many cases, key stakeholders will also be part of the planning process (e.g., a teacher on the planning committee who is also a representative for the teachers' union).

The basic rationale for sample size is the same as elsewhere: the size, complexity of organization, and within-grantee diversity will determine the sample size. In some cases, more stakeholders will be included. In others, we do not anticipate any respondents in this category who won't already have been selected based on another category. Overall, we anticipate an average of one, unique stakeholder per grantee.

As with other respondents, the lists will be initially generated by a review of grantee documents. For example, these typically reference if there is support from a business community organization (in which case we would request to interview the leader or TIF point person from the organization). Additional interviews might lead to us learning of changes in the grantee that require us to conduct an interview with a respondent not identified from initial documents. In that case, we would add respondents while keeping within our average of 10 respondents per grantee.

g. Are the three sites used to pilot the questionnaire excluded from the implementation study?

No. While we will review notes from earlier interviews to reduce respondent burden (in the rare case where respondents are the same—e.g., project directors who have retained their position), protocols changed (as is appropriate) during the piloting process (prior to submission of our original OMB package). Additionally, time has passed and the projects have presumably progressed since piloting. Finally, with such a small sample of respondents for any given grantee, we did not collect comprehensive data on these grantees. It would be harmful to the study to exclude them.

13. Questionnaires:

a. Please provide the final versions (see note in SS B4, which suggests that later versions are forthcoming) and please note specifically what has changed from the versions already submitted.

It is our understanding that our referencing of ED as a separate entity from the SRI-led research team made it unclear that the reviews and changes we describe were in fact reflected in the final protocols submitted with this package in June (Please see response to question 6). No later versions of the interview protocols have been developed since the submission of the OMB package. We have in a two cases made clarifications to the protocols as part of our development of researcher training processes. First, as described in part B of this question, we highlighted particular questions for which we only need the answer from one respondent. Second, we have indicated which questions ought to be asked primarily of the person in charge of each grantee's data system (in the case that grantees have such a position) with highlighting.

These clarifications should have no affect on interviewees' experiences with the evaluation; rather they ought to make it easier for researchers to conduct the interviews as designed. We have submitted the same final protocols we submitted before with highlighting (in case OMB might consider the highlighting to be a change).

b. We understand the potential utility of asking multiple actors their views on certain topics (e.g., how the program roll out has gone) but we don't understand why multiple participants need to be asked the same questions about factual matters, such as previous pay for performance initiatives in the state or target goal for numbers participating. Please explain.

We never planned to ask multiple actors the same questions about factual matters. As part of preparing our research team training, we highlighted the questions that need to be answered by only one, knowledgeable respondent. We believe this, in addition to our training, will largely eliminate the possibility of researchers asking unnecessarily redundant questions. (See Attached IC 3, IC4, IC 5, IC6.) As is indicated in some sections, some questions that are factual are asked across two or more protocols for purposes other than to gather factual data. For example, we ask about prior history of performance pay in the grantee to get a sense of how involved and aware the respondent typically is of performance pay in the locale. We ask about the design of the program to see how the program is perceived; we will rely on the project director only for an accurate description.