Impact Evaluation Design Plan for the HPOG 2.0 National Evaluation
Appendixes

National and Tribal Evaluation of the 2nd Generation of Health Profession Opportunity Grants (HPOG 2.0)

OPRE Report #2019-82

August 2019

HPOG
Health Profession Opportunity Grants
OFFICE OF FAMILY ASSISTANCE

National and Tribal Evaluation of the 2nd Generation of Health Profession Opportunity Grants (HPOG 2.0)

Impact Evaluation Design Plan for the HPOG 2.0 National Evaluation Appendixes

OPRE Report No. 2019-82

August 2019

Jacob Alex Klerman, David Judkins, and Gretchen Locke, Abt Associates

Submitted to:
Nicole Constance, Hilary Bruck, and Amelia Popham, Project Officers
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Contract No. HHSP233201500052C, Task Order HHSP3337016T

Project Director: Gretchen Locke
Abt Associates
6130 Executive Boulevard
Rockville, MD 20852

Contents

APPENDIX A: Topics in Impact Estimation

This appendix discusses six topics related to impact estimation. Section A.1 discusses covariate selection. Section A.2 discusses the use of post-randomization information in survey nonresponse adjustments. Section A.3 explains our plans for constructing regression-adjusted treatment and control group mean outcomes. Section A.4 expands on Section 3.2 of the main body of this plan on internal versus external inference, perhaps more aptly referred to as retrospective and prospective analysis. Section A.5 describes the method for estimating impact on a "pseudo-wage." Section A.6 further motivates the estimation of the differential impact of receipt of an hour of HPOG training versus an hour of non-HPOG training.

A.1 COVARIATE SELECTION

Section 2.3.1 proposed a tentative list of covariates to be included in impact regression equations and noted that we will defer final selection of covariates to the *Analysis Plan* corresponding to the *Short-Term Impact Report*. That list is rather long. Although covariates have the potential to increase precision, too many weak covariates can hurt precision.[1] Unpublished simulations show that the variance penalty increases as the ratio of non-significant to significant covariates grows.[2] Intuitively, if weak covariates are included in Equation 3.1, they will sometimes be out of balance and the regression adjustment will overreact to these non-meaningful aspects of the treatment/control sample balance.

To avoid this unintended variance inflation, we need to drop or otherwise reduce the influence of extraneous covariates. However, new research we have conducted shows that if we drop or downweight covariates based on simple analyses of the same data used in the evaluation, we will get biased estimates of the variance on the estimated treatment effects. This bias is negative, meaning that the variance estimates are too small. This in turn means that too many estimated impacts will be flagged as statistically significant.

In order to gain the variance reduction associated with controlling on true determinants of HPOG outcomes while still being able to estimate variances unbiasedly, the best procedure is to reduce the covariate list based strictly on information from an external data source with the same sets of available covariates and outcomes. In the case of HPOG 2.0, we have an almost perfect test bed in the data collected to evaluate HPOG 1.0. Although the collections of covariates and outcomes are not identical, there is substantial overlap, particularly in the covariates.

Using HPOG 1.0, we will develop two sets of covariates for HPOG 2.0. The first set of covariates will be used for outcomes that are not analyzed on a dataset that has been merged

---

1 This possibility appears to have been first noted by Tukey (1991). Despite sample sizes that were larger than will be obtained for most HPOG 2.0 programs (particularly when using the follow-up surveys), early analyses of PACE showed the pattern foretold by Tukey. Specifically, using a large number of covariates sometimes resulted in larger variances on estimated program effects than could be obtained with a smaller number of covariates. Liu (2011) derived explicit formulae for how much the variance increases when weak but strongly imbalanced covariates are included in regression models.

2 For example, with a sample size of 1,000, when there are three covariates that explain 57 percent of the variation of the outcome and 97 covariates that are not relevant to prediction of the outcome, the standard error of the effect of treatment is 11 percent higher with ordinary least squares than with a model that contains only the three relevant covariates (Austin Nichols, Abt Associates, unpublished simulations, 2016).

with National Directory of New Hires earnings data. The second set of covariates will be used for those outcomes that are analyzed on a dataset that has been merged with NDNH earnings data.

We will first identify the two collections of available covariates in the two analysis settings (OPRE server for NDNH analyses and Abt Associates server for survey analyses) and the sets of outcomes we intend to analyze in each analysis setting. Within each setting, we will use SAS/GLMSELECT to select a parsimonious model for each relevant.

More specifically, we will use the "LASSO" option with 10-fold "cross-validation" to pick the "optimal constraint." The LASSO stands for *least absolute shrinkage and selection operator*. With the LASSO, the sum of absolute values of the regression coefficients in a model is constrained to be less than a preselected value, the "constraint." If the constraint is small enough, many coefficients will be forced to zero. SAS/GLMSELECT estimates LASSO model coefficients with a wide range of values for the constraint. With 10-fold cross-validation, the sample is divided into 10 equal pieces. For each candidate value of the constraint, the LASSO is fit on a subsample in which one of the 10 pieces has been dropped and prediction errors are obtained for that same piece. Squared prediction errors across all 10 replicated are then calculated. The procedure picks the smallest value of the constraint that minimizes this cross-validated mean square prediction error. Whichever variables have nonzero coefficients in the model for that optimal constraint are selected.

Once we have a parsimonious covariate set for each outcome, we will then collate the collections. We anticipate that the count of covariates that appear in at least one outcome model will still be larger than desired. Despite the LASSO confirming that these variables have genuine predictive power, some of the coefficients are likely to be very small. We will further winnow the common set of covariates by eliminating those covariates for which the absolute standardized regression coefficient is less than 0.05. (A standardized regression coefficient of 0.05 means that an increase of one standard deviation in the baseline variable causes a change in the outcome of 0.05 standard deviation.)

A.2 Using Post-Randomization Information in Survey Nonresponse Adjustments

When presented with the idea of using post-randomization information from linked administrative data systems in nonresponse weighting, some researchers at first express concern. This concern arises from analogy with regression adjustment. It would never be acceptable to include post-randomization information as a covariate in a regression-adjusted estimate of a treatment impact.[3] We assert, however, that as long as two conditions are met, there should be no objections to using post-randomization information from linked administrative data systems in nonresponse weighting for survey-reported outcomes. The first condition is that the linked data are available for everyone in the study. The second condition is that all the steps in the creation of the weights use information only from sample members in the same random assignment group of the study.

In addition, given the challenges and complexities of linking, others have questioned whether such procedures are worthwhile even if valid.

This section addresses both of these issues.

[3] See the discussion of the use of contaminated baselines in Schochet (2008b).

A.2.1 Validity

We begin by addressing two concerns that might lead analysts not to use post-randomization information in nonresponse weighting: (1) concern that the procedure dilutes treatment-control contrasts; and (2) concern that the procedure facilitates *p*-hacking.[4] The argument is indirect, beginning with consideration of outcome variable scaling and imputation.

It is generally accepted practice in the conduct of experiments to recode multiple aspects of post-randomization experiences into scales and then to use these scales as outcomes, provided that the processes are defined before breaking the blind on the experiment.[5] This recoding process can include modeling of relationships among variables and using estimated model parameters to specify item weights in the scales. Imputation of survey-based outcomes based on linked administrative data can be viewed as a complex recoding of administrative outcomes and self-reported outcomes. Thus, such imputation should not be objectionable. Here it is, of course, critical that the imputation process never borrow information across random assignment groups (to avoid effect dilution) and that the imputation be carried out before preparation of any variables measuring treatment/control contrasts (to avoid *p*-hacking).

This is relevant to discussions of nonresponse weighting adjustments because every weighted mean of treatment/control respondents can also be expressed as an unweighted mean on the entire sample of randomized treatment/control cases where nonrespondents are imputed to have the value

(Eq. A.1) $$\tilde{y}_{Ni} = \frac{n}{n - n_R}\left(\frac{\sum_j w_{Rj} y_{Rj}}{\sum_j w_{Rj}} - \frac{\sum_j y_{Rj}}{n}\right)$$

where:

- the summations are on either the treatment or the control sample;
- n is the randomized sample size for the arm;
- n_R is the respondent sample size for the arm;
- w_{Rj} is the nonresponse-adjusted weight for the *j*-th respondent; and
- y_{Rj} is the reported value for the same respondent.

Thus, weighting based on post-random assignment outcome values is a form of imputation, which itself is a form of scaling, which everyone agrees raises no concerns about effect dilution or *p*-hacking. Thus, this type of weighting should also not be objectionable based on concerns about effect dilution or *p*-hacking.

---

4 "*p*-Hacking" is a term coined by Simmons, Nelson, and Simonsohn (2011) to describe a variety of analysis procedures that researchers may consciously or subconsciously employ to coax additional statistically significant results out of their data.

5 See, for example, Schochet (2008a), Guideline #3 and Appendix C.

A.2.2 Worthwhile Given the Challenges?

Because no variance reduction can be expected from nonresponse adjustment, the only reason to do it is if nonresponse biases are reduced. As discussed in Little and Rubin (2002), if nonresponse is ignorable—that is, has only minor consequences for measure outcomes—given baseline information, then unbiased estimates of effects can be prepared with nonresponse-adjusted weights that are only a function of that baseline information. However, if nonresponse is non-ignorable even controlling for baseline information, such estimates will be biased. So in considering whether to use linked post-randomization information to prepare nonresponse-adjusted weights, the critical question is whether the assumption of ignorability is more plausible given the union of baseline and linked post-randomization information than it is given just the baseline information.

This is an empirical question that has not been studied yet, partially because in many studies, biases from non-consent for linkage to administrative data might be as worrisome as survey nonresponse bias (Sakshaug & Kreuter, 2012). However, in HPOG 2.0, consent for linkage was a precondition for admission to the study, so unless there is a substantial and unexpected rate of consent withdrawals, there should be very little reason to be concerned about non-consent bias. Sakshaug and Kreuter (2012) estimated both biases and found that nonresponse biases were substantially larger than non-consent bias for some of the examined outcomes.

In the first wave of PACE reports, National Student Clearinghouse data were used in the construction of nonresponse-adjustment weights for survey data. These reports also evaluated the utility of these adjustments. Generally speaking, the utility of the weights appeared to be modest despite fairly large differences in response rates between the treatment and control groups in some programs. For the most part, unweighted regression adjustment on survey respondents based only on baseline data seemed to perform nearly as well as weighted regression adjustment, where the regression adjustment was based only on baseline data but the nonresponse adjustments were based on both baseline data and current NSC data.

For the second wave of reports, the PACE evaluation team will research whether this pattern continues to hold. Those results will be available before the first HPOG 2.0 *Analysis Plan* for the *Short-Term Impact Report* is drafted in 2019 and will help inform our decisions.

A.3 Constructing Regression-Adjusted Treatment and Control Group Mean Outcomes

As mentioned in Section 3.1.3, we will publish control and treatment group mean outcomes, but they will not be the simple weighted means. We will not show the simple means because the difference between them will not equal the estimated HPOG impact derived from a regression model. Instead, the means for control and treatment groups will be calculated as

$$\bar{Y}_C = \frac{\sum_i w_i(1-T_i)Y_i}{\sum_i w_i(1-T_i)}$$

$$\bar{Y}_T = \hat{\delta} + \bar{Y}_C = \hat{\delta} + \frac{\sum_i w_i(1-T_i)Y_i}{\sum_i w_i(1-T_i)}$$

(Eq. A.2)

where:

- $\bar{Y}_C$ is the actual mean of the outcome for the control group;
- $\bar{Y}_T$ is the reported "mean" of the outcome for the treatment group;
- w_i is a nonresponse-adjustment weight;
- $\hat{\delta}$ is the estimated impact; and
- *i* sums across members of the control group.

Note that both means are tabulated only on the control sample by virtue of the fact that $1-T_i=0$ on the treatment sample. The treatment mean represents the counterfactual projection of what would have happened if everyone in the control group had been allowed into HPOG.

Other common choices for publishing counterfactual projections are

(Eq. A.3)

$$\bar{Y}_T = \frac{\sum_i w_i T_i Y_i}{\sum_i w_i T_i}$$

$$\bar{Y}_C = \bar{Y}_T - \hat{\delta} = \frac{\sum_i w_i T_i Y_i}{\sum_i w_i T_i} - \hat{\delta}$$

and

(Eq. A.4)

$$\bar{Y}_T = \hat{\delta} + \frac{\sum_i w_i X_i \hat{\beta}}{\sum_i w_i}$$

$$\bar{Y}_C = \frac{\sum_i w_i X_i \hat{\beta}}{\sum_i w_i}$$

In Equation A.3, the treatment mean is actual (because *i* sums across members of the treatment group), and the control mean is a counterfactual projection of how the treatment group would have fared without HPOG. In Equation A.4, both means are counterfactual projections to the entire randomized population, because here *i* is summed across both the treatment and control groups. If the treatment and control samples are perfectly balanced on all the covariates in the regression model, then all three sets of means will be identical, but some chance imbalances are always expected.

The advantage of using Equation A.2 for the evaluation of training programs such as HPOG is that for zero-inflated outcomes[6] with high skew[7] where the program is expected to produce positive effects, it is possible for both $\frac{\sum_i w_i T_i Y_i}{\sum_i w_i T_i} - \hat{\delta}$ and $\frac{\sum_i w_i X_i \hat{\beta}}{\sum_i w_i}$ to be negative, which is awkward given that negative values are impossible. If, on the other hand, the program is expected to reduce the level of an undesirable outcome (e.g., poverty or food insecurity), then it would be preferable to use Equation A.3.

We will try to avoid this problem by constructing the outcomes in such a manner that expected impacts of HPOG are positive rather than negative. If the methodology nonetheless produces any negative estimates of counterfactual means, this is a signal of instability of the underlying models. Accordingly, we will remove all covariates from the model for that outcome.

A.4 Internal and External Inference

For HPOG 2.0, we will use two different inferential strategies on pooled samples to support the goals of powerful inference about HPOG 2.0 itself and provide useful guidance to funders and designers of future programs. As discussed in Section 3.2.3, the model for external inference—supporting future program design—will allow treatment effects to vary randomly from program to program for unknown reasons. There are some challenges in fitting such models when the sample sizes vary widely across programs and the total number of programs is modest. We expect that Judkins (2014), McNeish and Stapleton (2016a, 2016b), and Bloom, Raudenbush, Weiss, and Porter (in press) will have useful guidance on these issues, but that further research might be called for, particularly, how to deal with unstable estimates of the between-site variation in effects, for which Bayesian methods might offer useful improvements as in McNeish (2016).

For internal inference on NDNH analyses, we will use fixed intercepts for the programs (i.e., program dummy variables) and a single common slope across all programs. For internal inference on survey-based analyses with nonresponse adjustment, we will use survey regression software, treating the programs as both sampling strata and fixed effects on the right-hand side of the model statement.

A.5 Estimating Impact on "Pseudo-Wages"

In the evaluation of a training program's impact on participant earnings, it is of substantive interest whether increases in earnings (*E*) arise because of an increase in hours worked (*H*) or from higher (average) earnings per hour (i.e., a higher wage rate, *w*), where *E* = *H* * *w*. The classical justification for job training is that it increases "human capital" leading to higher (hourly) wages. Yet there is some evidence that earnings impacts instead come from faster reemployment and more hours worked once employed (Ashenfelter, 1978; Heckman & Smith, 1999). In any case, the relative importance of each component is informative about the pathway

6 Zero-inflated variables are characterized as non-negative variables that are continuously distributed above zero but have a non-negligible point-mass at zero. Prominent examples include training hours and earnings, as many people have zero training hours, whereas many others have zero earnings.

7 Skew is where a distribution has a longer tail to the right or left of the mean than on the other side.

through which any impact on earnings occurs. Thus, it is of substantive interest to decompose any earnings impacts into impacts on hours and impacts on wages.

For some purposes, it is useful to estimate the impact on "pseudo-wages," and we will do so. In brief, earnings and hours are both observed for everyone randomized, so we can estimate the effect of HPOG on them consistently using conventional methods. We define the pseudo-wage as the residual of change in earnings less change in hours. Then, because the percentage impact on earnings is (approximately) the sum of the percentage impacts on hours and on the pseudo-wage, if the relative effect of HPOG on earnings is larger than the relative effect on hours, then there must also be an impact on wages or on the composition of those working. As a substitute for the desired effect on wages among the always employed, we calculate impact on pseudo-wages.

We define the pseudo-wage, *p*, for a finite population as

(Eq. A.5) $$p \equiv \frac{\bar{E}}{\bar{H}}$$

where:

- $\bar{E}$ is average earnings for the population; and
- $\bar{H}$ is average hours worked for that population.

Thus, *p* is defined to be the value, such that its product with average hours is equal to average earnings.

We then define the impact of an intervention on *p*, $\hat{\delta}_p$, as the difference in pseudo-wages between treatment and control group participants:

(Eq. A.6) $$\hat{\delta}_p = p_T - p_C = \frac{\bar{E}_T}{\bar{H}_T} - \frac{\bar{E}_C}{\bar{H}_C}$$

We note that this impact is only meaningful at the population level and can only be computed in populations where at least some of the sample find work with or without treatment. All the other impacts discussed in this report are population averages of person-level impacts. We will caution readers not to interpret the effect of HPOG on the population pseudo-wage as saying anything at all about effects of HPOG on person-level wages for the always employed.

We can compute the standard deviation (and confidence intervals) of this impact estimate using the delta method. In particular, first note that $\mathrm{Var}\left(\hat{\delta}_p\right) = \mathrm{Var}\left(p_T\right) + \mathrm{Var}\left(p_C\right)$ because p_T and p_C are independent (one is estimated on the treatment group, the other on the control group). To calculate $Var(p_g)$ for either the treatment or control group ($g = T$ or C), where $p_g = \bar{E}_g / \bar{H}_g$, we first use the delta method to approximate $\bar{E}/\bar{H}$ around its mean μ_E/μ_H:

(Eq. A.7) $$\frac{E}{H} \approx \frac{\mu_E}{\mu_H} + \frac{1}{\mu_H}\left(E - \mu_E\right) - \frac{\mu_E}{\mu_H^2}$$

where we have dropped the subscript *g* for simplicity. The variance of E/H can then be estimated as

(Eq. A.8)

$$Var\left(\frac{\bar{E}}{\bar{H}}\right) \approx Var\left(\frac{\mu_E}{\mu_H} + \frac{1}{\mu_H}\left(\bar{E}-\mu_E\right) - \frac{\mu_E}{\mu_H^2}\left(\bar{H}-\mu_H\right)\right)$$
$$\approx \frac{1}{\mu_H^2} Var\left(\bar{E}\right) + \frac{\mu_E^2}{\mu_H^4} Var\left(\bar{H}\right) - 2\frac{\mu_E}{\mu_H^3} Cov\left(\bar{E},\bar{H}\right)$$

(Note that earnings *E* and hours *H* will be correlated.) Translating this into sample measures, we can calculate the variance of $\bar{E}/\bar{H}$ for each group as

(Eq. A.9)

$$Var\left(\frac{\bar{E}}{\bar{H}}\right) \approx \frac{1}{n}\left(\frac{1}{\bar{H}^2} s_E^2 + \frac{\bar{E}^2}{\bar{H}^4} s_H^2 - 2\frac{\bar{E}}{\bar{H}^3}\rho s_E s_H\right)$$

where:

- s_E and s_H are the standard deviation of earnings *E* and hours *H* in the group;
- ρ is the correlation between *E* and *H;* and
- *n* is the sample size (of the treatment or control group, respectively).

Pulling this all together, we can test for whether there is effect of HPOG on pseudo-wages by examining the following:

(Eq. A.10)

$$\frac{\left|\hat{\delta}_p\right|}{\sqrt{\text{var}(\hat{\delta}_p)}} \approx$$
$$\frac{\left|\frac{\bar{E}_T}{\bar{H}_T} - \frac{\bar{E}_C}{\bar{H}_C}\right|}{\sqrt{\frac{1}{n_T}\left(\frac{1}{\bar{H}_T^2} s_{ET}^2 + \frac{\bar{E}_T^2}{\bar{H}_T^4} s_{HT}^2 - 2\frac{\bar{E}_T}{\bar{H}_T^3}\rho_T s_{ET} s_{HT}\right) + \frac{1}{n_C}\left(\frac{1}{\bar{H}_C^2} s_{EC}^2 + \frac{\bar{E}_C^2}{\bar{H}_C^4} s_{HC}^2 - 2\frac{\bar{E}_C}{\bar{H}_C^3}\rho_C s_{EC} s_{HC}\right)}}$$

We are now in a position to calculate the share of the change in earnings due to the change in hours and the change in the pseudo-wage. To do so, we decompose the percentage change in earnings (*E*) into percentage change in hours worked (*H*) versus percentage change in the pseudo-wage (p_w). As is standard in models of this form, note that for small changes, the percentage change in earnings is approximately additive in the percentage changes of its underlying components: $\hat{\pi}_E \approx \hat{\pi}_H + \hat{\pi}_{p_w}$, where for each element, π is the percentage change of the impact, measured in terms of the control group mean (e.g., $\hat{\pi}_E = \hat{\delta}_E/\bar{E}_C$). We have

estimates for each of the impacts—$\hat{\delta}_E, \hat{\delta}_H, \hat{\delta}_{p_w}$—and their standard errors, as well as for each of their control group means—$\overline{E}_C, \overline{H}_C, \overline{p}_{wC}$.

We can thus divide the percentage impact on earnings into a percentage impact on hours and a percentage impact on the pseudo-wage, and compute standard errors for those claims (e.g., can we reject half?). Herr and Klerman (2017) provide a worked example.

A.6 Estimating the Difference in Impacts between HPOG and non-HPOG Training

Section 3.6 argued that we can solve for the differential impact of receipt of an hour of HPOG and an hour of non-HPOG training from data on two programs. This appendix shows explicitly how to do this under the assumption that one of the two programs has no non-HPOG training—neither for the treatment group nor for the control group (i.e., the HPOG program is the only available training program). Once the reader understands the algebra for this case, the algebra for the general case (when both programs have some non-HPOG training) should be clearer.

Suppose we observe a program *j* where those in the treatment group and those in the control group received no non-HPOG training. With this assumption, the second term in Equation 3.6 is zero.

(Eq. A.11)
$$\begin{aligned}\hat{\delta}_j &= \overline{y}_{j(d=1)} - \overline{y}_{j(d=0)} \\ &= \left\{\rho_{HPOG}\overline{H}^{HPOG}_{j(d=1)} + \rho_{non\text{-}HPOG}\overline{H}^{non\text{-}HPOG}_{j(d=1)}\right\} - \left\{\rho_{non\text{-}HPOG}\overline{H}^{non\text{-}HPOG}_{j(d=0)}\right\} \\ &= \rho_{HPOG}\overline{H}^{HPOG}_{j(d=1)} + \rho_{non\text{-}HPOG}\left(\overline{H}^{non\text{-}HPOG}_{j(d=1)} - \overline{H}^{non\text{-}HPOG}_{j(d=0)}\right) \\ &= \rho_{HPOG}\overline{H}^{HPOG}_{j(d=1)}\end{aligned}$$

where the first three lines are as in Equation 3.6, and the last line imposes the assumption (for clarity of argument, not required for estimation) that in this location, there is no non-HPOG training available.

Thus, for that program, we could estimate the impact of an hour of HPOG training, ρ_{HPOG}, as the estimated impact of the program divided by the average number of hours of HPOG training received (only in the control group):

(Eq. A.12)
$$\delta_j = \rho_{HPOG}\overline{H}^{HPOG}_{j(d=1)} \Rightarrow \rho_{HPOG} = \frac{\delta_j}{\overline{H}^{HPOG}_{j(d=1)}}$$

We can estimate the impact for a program, δ_j (e.g., by treatment/control differences or regression generalizations). We observe hours of HPOG training, $\overline{H}^{HPOG}_{j(d=1)}$. Thus, assuming there is no non-HPOG training at this location, we can estimate for the impact of another hour of HPOG training, ρ_{HPOG}.

Now consider a second location, *k*, where treatment and control group members received some non-HPOG training. We are trying to estimate the impact of non-HPOG training, $\rho_{non\text{-}HPOG}$. Solving Equation 3.6 (i.e., the third line of Equation A.11) for $\rho_{non\text{-}HPOG}$ yields:

(Eq. A.13) $$\rho_{non-HPOG} = \frac{\delta_k - \rho_{HPOG}\overline{H}^{HPOG}_{j(d=1)} + }{\left(\overline{H}^{non-HPOG}_{k(d=1)} - \overline{H}^{non-HPOG}_{k(d=0)}\right)}$$

We have estimates for everything on the right-hand side of Equation A.13. Again, we start with an estimate of the impact of being assigned to the treatment group in location *k*, δ_k (estimated as the treatment/control difference in the outcome, perhaps with regression adjustment). We observe average hours of HPOG and non-HPOG training for the treatment and control groups: $\overline{H}^{HPOG}_{k(d=1)}, \overline{H}^{non-HPOG}_{k(d=1)}, H^{non-HPOG}_{k(d=0)}$. Finally, from the location with no non-HPOG training, we have an estimate of the impact of another hour of HPOG training, ρ_{HPOG}, from the program with no non-HPOG training.

For the more general case of non-HPOG training at both locations, Equation 3.6 is a linear system of two equations (for the two location-specific impacts, δ_j, δ_k) and two unknowns (the returns to an hour of HPOG and an hour of non-HPOG training, $\rho_{HPOG}, \rho_{non-HPOG}$, respectively), and the other terms (hours of HPOG and non-HPOG training for treatment and control groups, $\overline{H}^{HPOG}_{j(d=1)}, \overline{H}^{non-HPOG}_{k(d=1)}, \overline{H}^{non-HPOG}_{k(d=0)}$). Standard results imply that we can solve for the objects of interest: the returns to an hour of HPOG and an hour of non-HPOG training $\rho_{HPOG}, \rho_{non-HPOG}$, respectively.

The discussion in this section can be interpreted as a proof of identification of the returns to an hour of HPOG and an hour of non-HPOG training, $\rho_{HPOG}, \rho_{non-HPOG}$, when we have two locations. In fact, we have more than two locations. As a result, we are over-identified; that is, there are multiple solutions corresponding to any arbitrary pair of locations. Instrumental variables methods implicitly choose the "best" estimate by averaging the estimates from each pair. Instrumental variables also show how to estimate the standard error of those estimated returns.

We leave further details to the *Analysis Plan* corresponding to the *Intermediate-Term Impact Report*, pending funding of this analysis. The depth of the analysis—including exploring various specifications and estimation methods—will depend on the level of funding.

APPENDIX B: Direct Assessment of Basic Skills

The logic models for career pathways programs typically include a path toward increasing earnings (Fein, 2012). The path identifies people with deficiencies in their basic skills and includes steps to remediate those deficiencies. After these skills are improved, people should be able to engage more effectively in occupational training, which should then lead to more and better employment with higher wages. Accordingly, it is important to determine experimentally if ACF-sponsored programs such as HPOG 2.0 improve the kinds of basic skills deemed most relevant for qualifying students to participate in education and training programs designed to prepare them for healthcare occupations.

Since telephone interviews with study subjects almost always play an important role in experimental evaluations of these programs, it would be useful to have a module that measures program participants' level of these basic skills over the phone. Although HPOG 2.0 grantees may conduct basic skills assessments at intake, these assessments cannot substitute for this role since they are not usually administered to members of the control group months after randomization. Further, the assessments the programs use are quite diverse across grantees, and the exact scores are generally missing in PAGES. To address this issue, our proposed approach has the potential to measure basic skills and thereby provide information about the adequacy of preparation of HPOG 2.0 participants for the early phases of their healthcare career training.

Several national and international surveys have been developed to assess adult numeracy and literacy, but almost all of these rely on face-to-face interviewing (an expensive mode to administer) or online administration (a mode infeasible for many evaluations that include low-income participants). Together with OPRE, the evaluation team at Abt Associates began exploring existing options. We contacted experts at Education Testing Services (ETS), Pearson Assessments, and ACT[8] to determine if they had—or could recommend—any existing instruments that would meet the study's needs. After several discussions, the team determined that these organizations did not have any suitable instruments.

The team did identify one phone-based assessment of verbal aptitude called the Computerized Adaptive Screening Test (CAST), used for prescreening of military recruits.[9] Tom Sticht—an expert in adult education, particularly oral and written language skills—has written favorably about it,[10] and an evaluation validated it for some uses.[11] The CAST appeared attractive for our purposes because it is designed specifically for telephone usage and only takes 10 to 15 minutes to administer. OPRE requested permission to use the CAST, but Defense Manpower Data Center declined the request.

Given the lack of available assessments that could be used in the evaluation of the impact of HPOG on basic skills, OPRE asked Abt Associates to create and pilot a module that would be

8 ACT is the organization that administers the college assessment test of the same name. The ACT test was formerly known as the American College Test, but is simply known now as the ACT.

9 The DMDC also is the home of the *Armed Services Vocational Aptitude Battery* (ASVAB), but the ASVAB requires in-person or on-line administration and lasts 50 minutes.

10 http://files.eric.ed.gov/fulltext/ED451383.pdf

11 http://www.dtic.mil/dtic/tr/fulltext/u2/a328971.pdf

similar to the CAST but meet the goals described above. Given the emphasis in many career pathways programs on contextualized remediation of basic skills and the fact that many programs, including HPOG, focus on teaching skills for healthcare occupations, we decided to focus on items with a healthcare "flavor." The idea is not to test healthcare knowledge *per se,* but rather to enable students in programs that use a health-care contextualized approach for basic skills instruction to have their achievements reflected in the verbal and mathematical scores.

Our primary purpose for the pilot is to identify a parsimonious subset of items that are relevant to success in healthcare training and that achieve the right mix of level of difficulty, as determined though item response modeling (IRT). Given this primary purpose, an assessment on which practicing medium-skill healthcare workers such as RNs would score 100 percent while people with basic skills much too low to enter occupational training would score near 0 percent, with a bell curve of scores for people between those extremes, would have the right level of difficulty. With the aid of grantees, we will recruit subjects who are nearly ready to start occupational training or who have started but not finished occupational training.

We developed a collection of 45 proposed items and are planning a small pilot test of that module before the start of interviewing for the HPOG 2.0 intermediate-term follow-up survey. The sample for the pilot will not include anyone sampled for survey follow-up. In addition to the assessment items, the pilot will contain some questions on education attainment and questions about use of basic and computer skills in everyday life. These questions are drawn from the Survey of Adult Skills (Programme for the International Assessment of Adult Competency, or PIAAC). These questions will afford the best opportunity for validation of the pilot module; that is, to supply indirect evidence that assessment scores correlate appropriately with the latent constructs of interest.

The evaluation team will analyze the results from the pilot and recommend a subset of items (12 vocabulary and 12 math items) for inclusion in a short basic skills assessment module. If the pilot is judged successful, the recommended basic skills module will be incorporated into the HPOG 2.0 intermediate-term follow-up survey.

To reduce the 45-item set, we will use factor analysis and item response theory (IRT) to identify a subset of the developed items that satisfy three goals:

1. Vary evenly in difficulty.
2. High reliability of a scale based on the smaller set of items.
3. High correlations of the reduced scale with earned credentials and with the PIAAC questionnaire items.

We will first discard any items that are either always or never answered correctly, and then we will use factor analysis to separate the items into sets that seem to be measuring different latent skills (presumably vocabulary and arithmetic skills) and item response theory to rank the difficulty of the items within each set. Given the need to keep the total administration time under ten minutes, we anticipate selecting about 12 vocabulary items and 12 arithmetic items. We expect each group of 12 will contain four items of (relatively) low difficulty, four items of (relatively) medium difficulty, and four items of (relatively) high difficulty. We defer the exact

groupings until analysis. We note that the low/medium/high range will be for the HPOG 2.0 population, few of whom have completed more than a high school education.

We anticipate selecting the items that best satisfy these goals using nonlinear programming. Specifically, we will aim to maximize the correlation of the reduced set with the full set for the latent skill, subject to the constraint that we have four easy items, four moderately difficult items, and four difficult items. After this, we will correlate the reduced set with earned credentials and the PIACC questionnaire items. If some excluded items have higher correlations with these validation scales than the tentative reduced scales, we will explore strategic substitutions. If timings indicate that the maximum number of items that can be administered within ten minutes is less than 16, then we will alter the number of items in each difficulty set. For example, we could have two easy, three moderate, and two difficult items for each of vocabulary and numeracy, giving a total of ten items to be administered to any single respondent. Alternatively, we could look to make cuts in other areas so that the assessment can last longer than ten minutes.