November 2023

Baseline Inequivalence and Matching

Both randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) can provide evidence of intervention impacts in Teen Pregnancy Prevention (TPP) effectiveness studies. However, for an RCT or a QED to convince a skeptical reader that the intervention caused the observed impact, the intervention and comparison groups in the impact analytic sample should be equivalent on key characteristics measured before the study began (that is, baseline characteristics) that influence outcomes.[1] In this brief, we discuss why baseline equivalence is important, how to assess it, and how to address baseline inequivalence, paying particular attention to meeting the requirements of the current Teen Pregnancy Prevention Evidence Review (TPPER).

Baseline inequivalence in impact evaluations and why it is a problem

In theory, the RCT and QED designs currently being used by TPP grantees to estimate intervention impacts can produce rigorous evidence of intervention effects, provided that the two groups being compared in either design are comparable at baseline on characteristics that influence the outcomes of interest.

In a well-executed RCT, the study sample is randomly divided into the intervention and comparison groups and, therefore, will be similar on all measured and unmeasured characteristics at baseline (any differences will be due to random sampling errors). As a result, any intervention-comparison group differences in outcomes can be attributed as the effect of the intervention.

In a well-executed QED, the intervention and comparison groups are not created by randomly dividing the study sample into the two groups. Although the intervention and comparison groups in a well-executed QED can be shown to be similar on key characteristics measured at baseline, there is a possibility that the two groups differ on unmeasured characteristics. Therefore, we are less confident (than we are with an RCT) that differences in outcomes between the two groups in a QED solely reflect the effect of the intervention—they may also reflect unmeasured differences between the two groups that affect outcomes. Therefore, the evidence from a QED is considered lower in quality than the evidence from a well-executed RCT.

In practice, two issues—which relate to study setup and sample loss—can render the study groups in both an RCT and a QED inequivalent and thus can undermine the strength of the evidence they produce.

The first issue is that problems with the **initial study setup** can either cause nonrandom allocation of the sample in an RCT or result in mismatched groups in a QED, as follows:

- **Nonrandom allocation of sample in an RCT.** In a random assignment study, if the process used to assign study participants to conditions is not effectively random, or the random assignment is

HHS Office of Population Affairs
Web: opa.hhs.gov | Email: opa@hhs.gov
LinkedIn: HHS Office of Population Affairs | YouTube: HHSOfficeofPopulationAffairs

undermined by intervention staff or participants, then there is no guarantee that the process will produce groups of youth that are equivalent on all measured and unmeasured variables. For example, intervention staff involved in random assignment might selectively choose participants who seem more willing to change their behavior to participate in the intervention condition (even if they are assigned to the comparison group), which invalidates the random assignment process. The sample members may actually be inequivalent on measured variables that are expected to influence outcomes, which could invalidate the assertion that any post-intervention differences in outcomes are attributable to the intervention.

- **Mismatched groups in a QED.** In a QED, if the initially assigned groups are drawn from substantially different populations, the groups may differ on key measured or unmeasured baseline characteristics that influence outcomes. In this case, any comparison of outcomes across conditions will produce a biased test of the intervention, due to the differences in the baseline characteristics of the groups. For example, if students drawn from a low-income area were offered the intervention, and students from a more affluent area served as the comparison group, the intervention-comparison group difference in outcomes would conflate intervention effects with the systematic differences in the composition of the intervention and comparison conditions.

A second issue that can create baseline inequivalence problems is **sample loss** as a result of nonresponse.[3] The intervention may affect whether or not an individual will participate throughout the study period and complete a follow-up assessment. For example, some intervention group members may drop out of a study soon after experiencing the program because they do not find the services useful. As a result, in both RCTs and QEDs in which the initially assigned groups were equivalent on key baseline variables, sample loss can produce final samples that are not comparable. Therefore, when outcomes are compared in the final samples (which will be *subsets* of the samples originally assigned to condition), the resulting impact estimates will be biased due to underlying differences between the intervention and comparison groups being used to estimate the impacts.

The analytic sample is the sample of youth with observed data for the outcome of interest at the point at which program intervention impacts are to be estimated. Establishing the equivalence of the intervention and comparison groups in this sample is necessary to convince skeptical readers that an impact estimate from this sample is credible.

There may be multiple analytic samples within a study if there are outcomes examined at several time periods.[2] In that case, baseline equivalence must be established for each analytic sample (corresponding to the follow-up period). If the sample sizes for two or more outcomes *within a specific follow-up* period vary slightly, it may be possible to construct a single analytic sample of youth who have complete data for all outcomes in that follow-up period for a simple, parsimonious presentation. Using that sample, you then would demonstrate equivalence between groups and estimate impacts for all outcomes in that follow-up period. If, however, there are substantially different response rates across outcomes within a follow-up period, you could consider creating two or more analytic samples for the follow-up period.

In the rest of this brief, we focus on assessing and establishing equivalence of the *analysis (or analytic) sample*. Establishing baseline equivalence on the analytic sample is necessary (according to TPPER) for

estimating credible program impacts. Doing so can mitigate concerns about baseline inequivalence threatening the internal validity of the study's evidence.

Baseline inequivalence in the analytic sample is a problem for TPP evaluations hoping to meet TPPER standards

Measured differences in outcomes between the intervention and comparison groups may result from the intervention's impacts, but may also be attributable to differences between the groups at baseline, before receipt of the intervention (that is, baseline inequivalence). TPPER standards recognize that baseline inequivalence can affect impacts, hence they stipulate that a study with substantive baseline inequivalence is at risk of being unable to convincingly demonstrate the effectiveness of the intervention and, therefore, receiving the lowest evidence rating. In particular, RCTs with high levels of sample attrition or QEDs with statistically significant intervention-comparison group differences on a key baseline measure can receive the lowest possible evidence rating, due to these threats to internal validity.[4] The rest of this brief describes the steps a researcher would take to demonstrate equivalent samples for studies with these problems, and to create equivalent samples if baseline inequivalence between groups is a concern.

Step 1. Deciding what variables to examine when assessing baseline equivalence

In general, to convince a skeptical reader that the intervention is solely responsible for the intervention-comparison group differences in outcomes, it is necessary to show that the two groups are equivalent on key baseline characteristics that are expected to influence the outcomes of interest. The TPPER standards have clear minimum requirements for demonstrating baseline equivalence:

We suggest assessing baseline equivalence on other key variables that are expected to influence outcomes, if such baseline data are available. For example, the study might assess equivalence on attitudes toward sex, knowledge about contraception and pregnancy, other measures of risky behavior (alcohol and drug use), or other variables that have been shown to correlate with sexual behaviors among teens. Since these variables are expected to influence outcomes, whenever possible, they should be examined for baseline equivalence to ensure that differences in these variables are not confounded with intervention impacts.

> *"... in order to receive the moderate study rating, quasi-experimental comparison group studies and random assignment studies with concerns about sample composition change are required to demonstrate that the intervention and comparison groups were similar at baseline on three key demographic characteristics: age or grade level, biological sex, and race/ethnicity. For studies with sample members at least 14 years old at baseline (or eighth grade or higher), the study authors must also establish baseline equivalence on at least one behavioral outcome measure (for example, rates of sexual initiation). This criterion is not applied to studies with younger sample members because rates of sexual risk behaviors are typically low for this age group."*
>
> — (TPPER Protocol version 7.1, p. 7-8)[5]

HHS Office of Population Affairs
Web: opa.hhs.gov | Email: opa@hhs.gov
LinkedIn: HHS Office of Population Affairs | YouTube: HHSOfficeofPopulationAffairs

Step 2. Documenting baseline equivalence of the analytic sample

Under version 7.0 of the TPPER protocol, baseline equivalence is determined by examining the magnitude of the difference in key characteristics across conditions. If the reported difference of a specified baseline characteristic is greater than 0.25 standard deviations in absolute value, the groups are considered to be non-equivalent. In addition, depending on the size of the baseline difference, a statistical adjustment may be required when estimating program effects, to produce a credible impact estimate.

- For demographic characteristics, when differences in the specified baseline characteristics are greater than 0.05 and lower or equal to 0.25 standard deviations, the analysis must include a statistical adjustment to meet the baseline equivalence requirement. Differences equal to 0.05 standard deviations or less require no statistical adjustment.[6]
- For baseline measures of the outcome, any difference of 0.25 standard deviations or less must be statistically adjusted for.

The first step in conducting the assessment is to create the analytic sample for a particular follow-up period of interest. As described earlier, this data set should initially contain those sample members who have valid assessment values for the follow-up period of interest. In addition, to show the baseline equivalence of that sample (that is, the sample in which you wish to compare outcomes), you should remove any sample members who do not have baseline assessments for the key variables described above. This will ensure that the ultimate analytic sample will have complete data for the outcome of interest, as well as all key baseline variables.[7]

The table shell below (Table 1) provides a template to use for your assessment. The final column indicates the statistical the difference between the intervention and comparison group means for each baseline characteristic of interest, in standard deviation unites (i.e., standardized). This number can be calculated by dividing the Mean difference (raw) by the pooled standard deviation of the characteristic (calculated by combining the standard deviation of the intervention and comparison group means). This is the key variable that the TPPER team examines when assessing baseline equivalence.

As shown in Table 1, you should document the sample sizes of the two groups in the analytic sample, *average* values of the continuous baseline measures (or the prevalence rates for dichotomous measures), and the standard deviations of the measures (if continuous). The intervention-comparison group difference in the average value of each measure should also be computed and tested to determine whether it is significantly different from zero. Importantly, the approach for conducting these statistical tests should be consistent with the study's design, so it may require taking into account clustering or stratification of the sample. For example, if the study randomly assigned schools to a condition within districts, the statistical test of baseline equivalence should incorporate dummy variables for districts and a clustering adjustment, such as school random effects or Huber-White clustering corrections for schools (Williams, 2000).

Although not currently required by TPPER, we recommend conducting an inferential test of the intervention-comparison group difference to ensure that the groups are not "statistically significantly different" even if the magnitude of the difference is small. While TPPER does not use this information in its assessment of the evidence, some readers (or journal referees) may want to use inferential statistics to understand comparability of the analytic sample at baseline.

Table 1. Analytic sample: Summary statistics of key baseline measures, for youth completing [survey name] as of [time stamp]

Baseline measure	Intervention group		Comparison group		Baseline differences		
	Mean (or %)	Standard deviation[a]	Mean (or %)	Standard deviation[a]	Mean difference (raw)	*p*-value of the difference	Mean difference (standar-dized)
Age or grade level							
Biological sex							
Race/ethnicity							
American Indian or Alaska Native							
Asian							
Native Hawaiian or Pacific Islander							
Black or African American							
White							
More than one race							
Unknown or not reported							
Behavioral measure, such as sexual initiation (for studies with youth at least 14 years old)							
Sample size							

Note: [Describe the analytic procedure used to test the intervention-comparison group difference in baseline means – for example, a t-test or random effects analysis.]

[a] Include if a continuous measure.

Best practices for continuous variables assessed at baseline

When assessing baseline equivalence of *continuous* measures, researchers should consider whether alternate specifications for those measures should be examined. This goes beyond the minimum mean difference requirements in the TPPER standards but may capture important differences that could confound intervention impacts. For example, suppose a study that includes girls ages 12 through 18 looks at an outcome of teen pregnancy.

TPPER standards require that the *average* ages of the intervention and comparison groups are not significantly different from each other. However, teen pregnancy can vary markedly in particular age categories, such as those under 14, ages 14 to 16, and over 16. Examining only *average* ages of this hypothetical study's intervention and comparison groups could miss important differences between the two groups in their age *distributions* that could be confounded with intervention impacts. Given this possibility, it would be useful for researchers conducting this study to also compare the age distributions of the intervention and comparison groups, in addition to assessing differences in averages in key measures at baseline. This could be done by separately examining each age category as a dichotomous variable in the baseline equivalence assessment, and estimating a linear probability model to assess the difference in the prevalence rates of the age categories across intervention and comparison groups.

Best practices for binary and categorical variables assessed at baseline

When assessing baseline equivalence of *binary and categorical* measures, researchers should consider whether it is important to examine combinations of the measures in addition to examining them individually, to address questions of intersectionality. Like the previous example, the TPPER standards do not require this type of analysis, but such an assessment might increase the face validity of the results.

For example, suppose a study includes girls who are all 16 years old.

TPPER standards require that race/ethnicity and at least one behavioral measure, such as sexual initiation, of the intervention and comparison groups are not significantly different. However, examining intervention-comparison group differences in race/ethnicity *separately* from group differences in sexual initiation could miss important differences between the two groups in the *combinations* of these measures. Even if the sexual initiation rates and race/ethnicity profiles of participants look similar across the intervention and comparison groups, the prevalence of the various combinations of sexual initiation and race/ ethnicity may differ across the intervention and comparison groups. To ensure that intervention-comparison group differences in combinations of race/ethnicity and sexual initiation are not an issue, researchers conducting this hypothetical study could examine group differences among individuals with the various combinations of these measures. This could be done by conducting the baseline equivalence assessments for combinations of variables, as shown in the table shell below (Table 2) for sexual initiation and race/ethnicity.

Table 2. Analytic sample: Summary statistics of combinations of key baseline measures for youth completing [survey name] as of [time stamp]

Baseline measure	Intervention group	Comparison group	Differences		
	Percentage	Percentage	Percentage points	*Effect size*	*p*-value of difference
White non-Hispanic and previously sexually initiated					
White non-Hispanic and not previously sexually initiated					
Hispanic or other race and previously sexually initiated					
Hispanic or other race and not previously sexually initiated					
Sample size					

Notes: [Describe the analytic procedure used to test the intervention-comparison group difference in baseline percentages.]

Step 3. Improving baseline equivalence in the analytic sample to potentially meet TPPER standards

If the diagnostic procedure outlined above reveals evidence of inequivalence, it is necessary to revisit what is considered the analytic sample for estimating intervention impacts. There are a number of equating approaches that use sample trimming, matching or weighting that can be used to attempt to mitigate baseline inequivalence in an analytic sample. (The “equated sample” is an analytic sample created using an equating process such as exact matching, propensity score matching, or a weighting approach.)

Exact matching method

A straightforward way to implement a potential matching approach is to select, for each intervention group member, a comparison group member who is *identical* on each characteristic of interest. For example, to identify groups that are equivalent in age, biological sex, and race/ethnicity, select for each intervention group member a comparison group member who has the same values for the characteristics. Because the initial comparison group was baseline inequivalent with the intervention group, this approach will yield a subset of the comparison group ultimately included in the impact analysis. Identifying an exact match for each intervention group member should ensure that the analytic sample produced by the matching procedure meets the TPPER requirements for baseline equivalence—in fact, this exact matching will ensure that the two groups have identical characteristics. This approach can be especially useful if a handful of outliers produce large group differences, and removing those outliers makes the intervention and comparison groups largely identical.

A potential limitation of this approach is that the comparison group may not contain an exact match for every intervention group member on all the characteristics that must be equivalent at baseline. For example, if four dichotomous variables are used to select a comparison group (such as dummy variables for age, biological sex, race, and ethnicity), there are 16 possible combinations of values for those

variables. The combination of characteristics for some intervention group members may not exist in the comparison group, and therefore, the number of exact matches might be a small subset of the original sample; this will limit power for the subsequent impact estimates.

Propensity score methods

Researchers who encounter problems using exact matching approaches should consider an alternate option: using propensity scores. This approach uses analytic methods to identify a subset of the original comparison group that is similar to participants, *on average*. The process generally involves two steps: (1) calculating a propensity score—a single number that can be used to assess the similarity between individuals on multiple measures—for each intervention and comparison group member; and (2) selecting the subset of comparison group members whose propensity scores are similar to those of intervention participants. This approach does not necessarily identify for each participant an exact match from the comparison group (as the first approach does), but it can identify a subset of comparison group members who are similar to participants *on average*. The Technical Appendix provides more details about using propensity scores to identify a comparison group that is, on average, similar to the intervention group along all key variables examined in the evidence review.

Weighting methods

An alternative to using propensity scores and matching sample members based on the propensity score is to calculate weights and using those weights to produce more credible impact analyses. For example, authors occasionally estimate inverse-propensity weights or entropy-balancing weights (Hainmueller, 2012), and use those weights in both the estimation of program impacts and the demonstration of baseline equivalence for the analytic sample.

General TPPER requirements related to equating approaches

The current TPPER standards outlines additional requirements for studies that use equating approaches:

- Equating approaches must only include exogenous variables
- Success of equating will be assessed by comparing the baseline differences in the matched or weighted sample, as described earlier
- Adjusting for the propensity (or other equating score) is insufficient as a statistical adjustment by itself
- If a study uses weighting approaches, it must document that the sum of the weights is less than or equal to the number of observations in the analytic sample

Step 4. Estimating impacts based on an equated sample

To estimate impacts from an equated sample, researchers should take certain steps to ensure that the impact estimation is likely to meet TPPER standards.

As stated above, before estimating intervention impacts, the baseline equivalence of the equated sample should be demonstrated and shown, using a format such as the one shown in Table 1. By demonstrating that this equated sample is equivalent at baseline, the study will have the potential to meet TPPER standards (provided that the impact analysis is conducted appropriately).

To ensure that impact estimates are credible and likely to meet TPPER standards, researchers should:

1. Estimate impacts using the sample of intervention and comparison group members who are baseline equivalent along the pre-intervention measures mentioned above.
2. Adjust the impacts for the pre-intervention measures that require a statistical adjustment, as well as variables that were included in the equating analysis. We also suggest adjusting impacts for any other pre-intervention measures that are correlated with outcomes. (This will improve precision of the impact estimate and adjust for any other differences between conditions).
3. Conduct a statistical test of the significance of the impact estimates that reflects the study's design:
   - For clustered designs—such as those that randomly assigned centers to intervention or comparison status but analyzed outcomes of individuals—ensure that the statistical tests account for the clustering of individuals in the groups.
   - For designs that conducted intervention-comparison assignment within strata or blocks, account for the number of strata created when conducting the statistical tests (for example by including dummy indicators for each stratum as a covariate in the impact analysis).
4. If equating approaches were used to identify the ultimate analytic sample, test whether the impacts are sensitive to alternate approaches used to generate the baseline equivalent sample. See the Technical Appendix for details on the types of robustness and sensitivity assessments that should be examined in this situation.

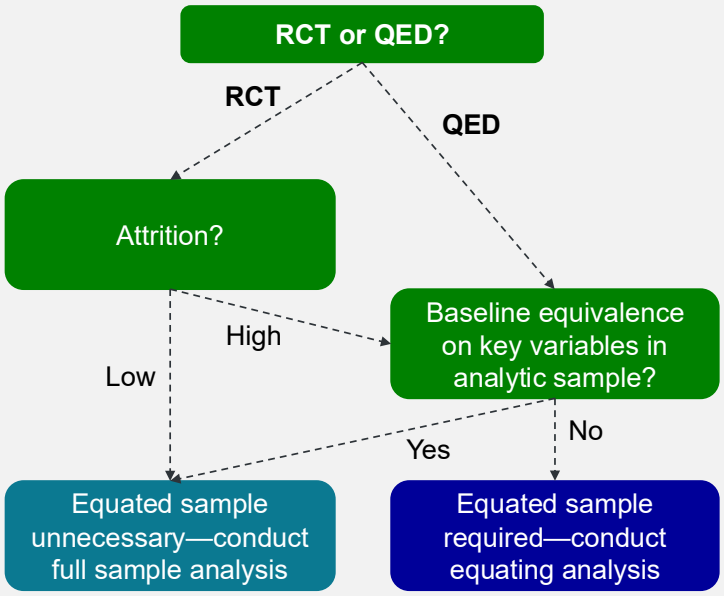
To present results from these analyses, consider showing impact results in a table, such as Table 3 below.

Step 5. Documenting matching results in a paper or final report to align with best practices

In TPP final evaluation reports, the approach for presenting impact estimates based on matching to improve baseline equivalence will depend on the study design (RCT or QED), level of sample attrition (for RCTs), and the degree to which the analytic sample groups are equivalent on key variables at baseline.

The following flow chart (Figure 1) illustrates the logic that should inform your approach. There are two scenarios that will require a matching analysis in order to be eligible for a moderate evidence rating (rather than a low rating):

Figure 1. Decision rules to inform if matching analysis is necessary



1. RCTs with high levels of sample attrition and a lack of equivalence on a key characteristic at baseline for the analytic sample.[8]
2. QEDs with a lack of equivalence on a key characteristic at baseline for the analytic sample.

In all other scenarios, the study will not be perceived to have a baseline equivalence issue, and therefore, an equating analysis is unnecessary.

Table 3. Post-intervention outcome measures and effects for analytic sample youth completing [survey name] as of [time stamp]

Outcome measures	Intervention group		Comparison group		Estimated effects		
	Mean (or proportion)	Standard deviation[a]	Mean (or proportion	Standard deviation[a]	Mean difference (raw)	Effect size difference	*p*-value of difference
Measure 1							
Measure n							
Sample size							

Note: [Describe the analytic approach used here, to align with the design and with the analytic approach used to demonstrate equivalence.]

[a] Include if a continuous measure.

References

Austin P. “An introduction to propensity score methods for reducing the effects of confounding in observational studies.” *Multivariate Behavioral Research*, vol 46, 2011, pp. 399–424.

Campbell, D., and J. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand-McNally, 1963.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis*, 20(1), 25-46. doi:10.1093/pan/mpr025

Rosenbaum P. R. “Model-based direct adjustment.” *The Journal of the American Statistician*, vol 82, 1987, pp. 387–394.

Rosenbaum, P. R. *Observational Studies*. New York: Springer, 2002.

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

Williams, R. L. “A note on robust variance estimation for cluster- correlated data.” *Biometrics*, vol. 56, 2000, pp. 645–646.

This brief was written by Russell Cole and Roberto Agodini from Mathematica for the HHS Office of Population Affairs under contracts #HHSP233201300416G and HHSP233201500035I/75P00122F37068.

TECHNICAL APPENDIX: Estimating and Using Propensity Scores to Obtain a Baseline-Equivalent Sample

A propensity score $\lambda(x)$ represents the probability of receiving the intervention (T = 1), given a set of characteristics *x* (Rosenbaum and Rubin 1983; Rosenbaum 2002). More formally,

$$\lambda(x) = \Pr(T = 1 \mid x),$$

where *x* includes key baseline characteristics that are expected to be related to intervention status and outcomes.

In the intervention evaluation literature, this propensity score can be used to produce an unbiased impact estimate, under the assumption that all important covariates are observed in *x*. Specifically, Rosenbaum and Rubin showed that statistical matching using propensity scores can be used to select a subset of the comparison group that is similar, on *average*, to intervention participants along those characteristics, which can facilitate the generation of an internally valid impact estimate.

The following general steps outline how to use propensity scores to identify a subset of the comparison group that is similar to intervention participants for the estimation of an internally valid impact. These steps are described in additional detail below.

First, use intervention participants and all potential comparison group members to determine how each baseline characteristic that affects outcomes also affects intervention participant status. Then, using this information, assign to each intervention participant and each potential comparison group member a propensity score that summarizes how each individual's baseline characteristics collectively influence intervention participant status. Finally, select a subset of the comparison group whose propensity scores are similar to those of intervention participants. This subset of comparison group members with propensity scores similar to those of the intervention group will allow for a more credible, internally valid estimate of intervention impacts than one based on a larger sample of comparison group members that is baseline inequivalent.

More specifically, the following steps can be used to assign a propensity score to each intervention participant and potential comparison group member:

1. Code an indicator variable equal to one for each intervention participant and zero for each individual in the pool of potential comparison group members. Call this indicator variable P.
2. Define indicator and continuous variables that represent the demographics and preintervention outcomes of intervention participants and potential comparison group members. For the purposes of producing an internally valid comparison that can meet TPPER standards, these variables should include demographics (biological sex, race/ethnicity, age), behavioral assessments of the outcomes of interest measured at baseline (if applicable), as well as other variables measured at baseline that

are expected to influence intervention assignment as well as follow-up outcomes of interest. Call this collection of variables X.

3. Using intervention participants and potential comparison group members, estimate a probability model—such as a logit or probit—where the dependent variable is P and the independent variables are X. Although some might argue that it is necessary for the probability model to align with the study design, we advocate for using a simple approach that does not take into account clustering or stratification, regardless of design:
   - **Clustered design.** Since the purpose of the propensity modeling approach is to obtain the correct parameter estimates for producing propensity scores, and not adjusting the standard errors of the parameter estimates, we do not feel that it is necessary to move away from a standard logit or probit regression approach in order to obtain plausible parameter estimates. As such, we suggest ignoring clustering in the estimation of the propensity scores.
   - **Stratified design.** Since the TPPER focuses only on the demonstration of baseline equivalence on a subset of variables (which we have suggested including in the propensity model), it is unnecessarily restrictive to conduct propensity modeling separately by strata, since the end result of the modeling and matching procedure (described below) can produce groups that are equivalent on the key characteristics of interest. As such, we suggest ignoring strata in the estimation of the propensity scores to increase the ease of estimation and likelihood of identifying matches for each participant.

   Results from the probability model will include parameter estimates, or a collection of values that indicate how each respective X affects P. Call this collection of values β .

4. For each participant and potential comparison group member, define a variable that equals the predicted probability of treatment (this will be the transformation of the sum of each β value times each respective X value and can be requested as an output in standard statistical packages). Call this variable P*. P* equals each individual's propensity score.

5. Select the subset of the comparison group for analysis using P*.[9] First, identify the subset of comparison group members whose propensity score falls within the minimum and maximum values of intervention participants (known as, the region of common support). Then, for each intervention group member, select a single comparison group member to serve as a potential match. The general approach for matching is to identify comparison group sample members with propensity scores that are very close to the propensity scores of each intervention group member. There are a number of ways of identifying matches (for example, see Austin 2011 for a comprehensive listing of methods). Matching can be performed to minimize the total difference in propensity scores across all intervention members and their matched comparison group (optimal matching), or it can be performed to only allow matches of a certain quality to occur (caliper matching—in which matches are only considered if the propensity scores differ by less than a certain level, known as the caliper). Matching can be conducted with or without replacement (so that a comparison group member may be

matched to multiple members of the intervention group). Selecting *with* replacement is particularly important if there are few comparison group members who are similar to intervention participants.[10]

6. Assess baseline equivalence of intervention participants and the subset of comparison group members who are matched. That is, complete Table 1 in the main brief text and check for any intervention-comparison group differences on the pre-intervention measures.
7. If the two groups differ on a pre-intervention measure of interest, revise the propensity model (the logit or probit) used in step 3 above to include higher-order terms for continuous measures and/or interactions for binary/categorical measures that are significantly different from each other. That is, if variable X_1 is significantly different across groups, then re-estimate the propensity model to include higher-order versions of X_1 or interact X_1 with other variables that are strongly related to intervention status.
8. Stop when you have identified a subset of the comparison group that is baseline equivalent with intervention participants.

   Rosenbaum and Rubin (1983) showed that a comparison group selected using propensity scores can produce unbiased impact estimates if two conditions are satisfied: (1) all the characteristics that are related to participant status and outcomes are observed, and (2) intervention and comparison group members with similar propensity scores are similar on individual characteristics. It is not possible to know for certain whether condition 1 is verified. However, adhering to the TPPER standards will ensure that several characteristics that the literature indicates are related to the outcomes of interest are included in the analytic model, and researchers can use additional data appropriate for their own populations to supplement the analysis to further support this claim. Condition 2 can be verified through the iterative process of estimating a propensity model, identifying matches, assessing equivalence, and re-specifying the model as necessary.

   As the process for using propensity scores above demonstrates, at certain points in the process researchers may need to make a subjective decision. For example, researchers will need to decide what types of matching techniques they will use, and make additional decisions within each technique. In addition, if intervention participants and the subset of comparison group members selected with propensity scores differ on a pre-intervention measure, the propensity model will need to be revised until a balanced com- parison group is identified. It is possible that more than one way of revising the probability model will produce a comparison group that is baseline equivalent with intervention participants.

   Given that using propensity scores to obtain baseline-equivalent groups requires making a number of decisions, researchers should calculate impacts based on at least two versions of the analysis (using different matching approaches or using different specifications of the propensity model) to assess whether impacts are sensitive to the subjective decisions made by the researcher. If the impacts are sensitive, this should be mentioned when reporting the results. If the impacts are not sensitive to the researcher's decisions, then it is sufficient to provide a footnote in the results about the additional analyses that were conducted, and indicate that the results were substantively the same.

Endnotes

[1] This is known as the "selection" internal validity threat, as defined by Campbell and Stanley (1963).

[2] In studies with more than two conditions (e.g., three groups were randomly assigned to intervention 1, intervention 2, or no services), the steps laid out in this brief should be conducted separately for each contrast between groups analyzed in a final report.

[3] Nonresponse in the context of an RCT includes the loss of any sample members who were initially randomized but were not included in the ultimate impact analysis. Common sources of nonresponse in TPP Evaluations include non-consent (after random assignment), program dropout, and nonresponse at the focal follow-up period used to estimate intervention impacts.

[4] A brief on sample attrition available at RHNTC provides more information on how studies can assess this threat and determine whether a matching analysis is necessary to meet TPPER standards.

[5] Studies can meet TPPER standards if they meet these and other conditions laid out in the TPPER protocol (Version 7.1).

[6] Examination of baseline equivalence for demographic characteristics with multiple categories, for example race and ethnicity, can be done with the modal category.

[7] The TPP Evaluation Technical Assistance team described several missing data approaches in the https://opa.hhs.gov/sites/default/files/2020-07/copingwithmissingdata.pdf, that are appropriate for only RCTs with low attrition. For high-attrition RCTs and QEDs, the TPPER will require a demonstration of baseline equivalence of the analytic sample *without imputation*, and therefore, the analytic sample must include cases with complete records on all key baseline and outcome variables.

[8] In addition, cluster randomized trials that include sample members in the impact analysis who were not included in the sample at the time of random assignment (in other words, they joined the sample after random assignment) may also be required to demonstrate baseline equivalence of the analytic sample to be eligible for the moderate study rating. This requirement is enforced in contexts where the unit of assignment could potentially be exploited by joiners (for example, when classrooms within a school are the unit of assignment and a student may join a particular classroom in order to get the intervention).

[9] There are a number of options for identifying a subset of the comparison group that may be baseline equivalent to intervention participants. This brief focuses on the use of propensity matching approaches for obtaining equivalence of the analytic sample because matching approaches are straightforward, understandable to a broad audience, and will achieve the goal of improving the equivalence of the analytic sample. For information on alternate approaches (such as inverse weighting or stratification of the propensity score), see Rosenbaum, 1987, or Rosenbaum and Rubin, 1983.

[10] We recommend that matching occur with replacement, so that each intervention member can be matched to the comparison group member with the closest P* value – that is, each participant's optimal match, which ultimately produces the optimal match for the entire intervention group. That said, we are not advocating for the duplication of comparison group members in the ultimate impact analysis. Rather, a comparison group member who ends up matching to multiple intervention group members should only contribute a single observation (with weight equal to the other sample members in the analytic sample).