October 2023

Considerations for TPP Evaluations with Multiple Follow-Up Assessments

Research studies for teen pregnancy prevention (TPP) programs—such as randomized controlled trials (RCTs) and quasi-experimental designs (QEDs)—often include multiple follow-up assessments designed to test the differences between treatment and control (or comparison) groups over time. For instance, a study might have a baseline survey and follow-up surveys at both six and 12 months after random assignment. These two follow-up assessment points would enable researchers and evaluators to understand how much a program's effects vary over time.

The purpose of this brief is to provide guidance for analyzing TPP impact evaluations that have more than one follow-up period. The brief is designed for evaluators who have some familiarity with statistical modeling and with the Teen Pregnancy Prevention Evidence Review (TPPER) standards (v. 6; HHS 2022). Additional details and references for audiences interested in a deeper technical understanding are provided in footnotes throughout the brief and in a technical appendix.

This brief begins by discussing four decisions evaluators have to wrestle with at four different points in the process of assessing a program's impacts over time. They help researchers and evaluators determine the design of an analytic plan that is appropriate and informative for a given study:

1. Accounting for the baseline measure of the outcome
2. Choosing between separate or longitudinal models across follow-up times
3. Treating time as discrete or continuous in longitudinal models
4. Accounting for within-person correlation in longitudinal models

We follow this by discussing some other considerations that can affect the validity of impact estimates in longitudinal analyses.

Note: This brief focuses on individual-level RCTs (and intent-to-treat analytic approaches) with continuous outcomes. This is a more straightforward way to make the key takeaway points about conducting impact analyses with multiple follow-up points. However, the principles discussed here can extend to other study designs (such as clustered designs and QEDs) and apply to dichotomous or other outcome types.

Four analytic decision points in assessing impacts at multiple follow-ups

Decision 1: Adjusting for the baseline measure of the outcome

The first decision confronting researchers and evaluators is how to adjust for underlying baseline differences between the treatment and control groups, particularly with respect to the baseline measure of

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

the outcome of interest.[1] If the analysis does not adequately adjust for baseline differences in the outcome measure, the differences could mistakenly be interpreted as an impact of the intervention.

We discuss two approaches to adjust for baseline measures: (1) subtracting the baseline measure from the (expected) follow-up outcomes and (2) controlling for the baseline measure as a covariate.

Subtracting the baseline measure from the follow-up outcome measures

The first approach to adjusting for baseline differences between treatment and control groups is to subtract the baseline measure from the follow-up measures, which results in a quantity commonly known as a gain (or change) score. In a gain score analysis, the gain score is regressed on the treatment indicator. The coefficient for the treatment indicator, interpreted as the difference in gain scores between the treatment and control group, is the impact of the intervention.

A similar approach to the gain score model, and one seen more commonly in TPP research (particularly in studies with multiple follow-up periods), is a difference-in-differences (DID) analysis. A DID analysis effectively treats the baseline measure as an outcome period in the regression model. The outcomes at all time points (including the baseline) are regressed on an indicator for the treatment condition, an indicator that the observation is in the post-period, and a treatment-by-post period interaction. The coefficient for the interaction term is the difference between the pre-post difference (gain score) of the treated group and the control group. A researcher can translate this gain score in the context of their study. For example, a researcher might calculate a gain score that represents a 5 percentage-point difference between the treatment and control group in the changes in rates of risky sex. The DID approach can be extended to noncontinuous outcomes (for example, binary, categorical, or count outcomes) through a generalized linear model (logistic or Poisson regression).[2]

The DID approach is always a longitudinal model (includes more than one observation per individual). We delve into longitudinal models in discussing Decision 2. Importantly, in any longitudinal model, the estimated standard errors need to account for the fact that some of the residuals are from the same individuals and are thus correlated. We say more about how to appropriately calculate standard errors as part of Decision 4.

Controlling for the baseline as a covariate

An alternative approach to adjusting for baseline measures is to treat them as covariates in the regression analysis of follow-up scores. The model includes the baseline score in the same way statistical adjustment includes any other covariate variables.[3] This approach is sometimes known as analysis of covariance (ANCOVA). The treatment effect coefficient is interpreted as the difference in outcome

[1] Although we discuss this in the context of studies with multiple follow-up assessments, this is a consideration for all studies to address, including those with only one follow-up assessment.

[2] DID makes a key assumption known as *parallel trends*. The parallel trends assumption states that, in the absence of the intervention, the individuals who received the treatment would have experienced a change in outcome similar to what the control group experienced, on average. This assumption enables the researcher to attribute the average difference in gain scores as the impact of the intervention.

[3] See Lieberman and Su (2012) for an illustration of this approach in practice.

between individuals in the treated group and those in the control group, holding the baseline measure (and all other covariates) constant.

Both the DID and ANCOVA approaches have the potential to meet TPPER standards for adjusting for differences in baseline measures. In fact, they will often result in similar impact estimates, specifically when the coefficient on the lagged outcome is close to zero or when the treatment and control groups are well balanced on the measure at baseline. They can differ, however, and because the two approaches make fundamentally different assumptions about the relationship between baseline and follow-up measures, the choice between them should be based on a substantive understanding of this relationship (Imbens and Wooldridge 2009). For a QED or high attrition RCT to achieve at least a moderate rating, TPPER requires a statistical adjustment for the baseline measure of the outcome of interest, and an ANCOVA approach to baseline adjustment is considered a valid statistical adjustment. If there is a strong relationship between the baseline and follow-up measures (for example, a correlation coefficient [$r > .60$]), using the DID approach is also appropriate as a means to satisfy the statistical adjustment requirement of TPPER.[4]

Table 1 summarizes considerations involved in deciding how to adjust for baseline measures.

Table 1. Decision 1 considerations

Decision	Things to consider
1. How to account for the baseline measure of the outcome	Adjusting for baseline scores on the left-hand side of the model equation (DID approach) requires the parallel trends assumption to be valid to yield credible impact estimates.
	Adjusting for baseline scores on the right-hand side of the model equation (ANCOVA approach) is a more traditional statistical adjustment. This adjustment treats baseline scores as a covariate along with any other covariates specified for the analysis (for example, participants' characteristics).
	Both approaches are valid statistical adjustments that can potentially yield credible impacts that separate underlying baseline differences from the observed impact estimates.

Decision 2: Choosing between separate or longitudinal models across follow-up times

The second decision researchers and evaluators have to consider is whether to estimate impacts as separate models for each follow-up period or as part of a single longitudinal model.

Separate models for each follow-up assessment

Estimating impacts separately for each follow-up period can answer questions about program effectiveness at each follow-up period independent of information about program effectiveness across other periods. This approach is straightforward to implement using standard linear or generalized linear

[4] Because dichotomous outcomes rarely have a pre-post correlation coefficient that exceeds $r = 0.60$, we recommend *not* using a DID approach to adjust for baseline measures of dichotomous outcomes. The What Works Clearinghouse (WWC 2022) recommends the $r = 0.60$ threshold.

models. Moreover, it makes it easy to include different covariates for models examining impacts at different follow-up time points if there is a reasonable justification for doing so.

Nonetheless, modeling each follow-up assessment separately has two main limitations. The first is that the resulting estimates will rely on smaller sample sizes than those from a longitudinal model that includes all assessment time points. The result is they will be noisier (have more variance), and there will be lower statistical power to detect a statistically significant effect if one truly exists. Second, differential attrition between study groups can lead to a more biased estimate of program effectiveness when estimating impacts separately than would result from estimating impacts in a single longitudinal model. Some longitudinal models offer built-in protections from this risk of bias due to missing data under certain common scenarios. This is covered later in the brief, in the section on "threats to validity with repeated measures studies."

A single longitudinal model for all assessments

Longitudinal approaches include all assessment periods in a single regression model. Pooling outcome data in a combined analysis can yield more precise impact estimates and more powerful significance tests, particularly when the covariates (for instance, demographics or baseline measures of the outcome) in the model have a similar relationship with the different follow-up periods.[5] For example, if age is a covariate in a model and correlates positively with sexual behavior outcomes at all follow-up periods, the reduction in error accounted for by the covariate across all time points would result in increased statistical power.[6]

In addition, pooling data from all time points in a single longitudinal model can give a researcher more insight into the persistence of impacts over time than using separate models for each follow-up assessment would. By using multiple follow-up impact estimates in the same model, it is possible to examine any variation in estimates across time. This allows for simple inferential tests to assess whether impacts at different follow-up periods differ statistically from each other.

Furthermore, some would argue that analyses of longitudinal models allow for more credible comparison of these impact estimates than separate analyses at each time point when the composition of the samples varies across time periods (potentially due to sample attrition). Their argument would be that the longitudinal analyses will yield unbiased impacts in more scenarios than separate analyses for each time point would.[7]

One potential limitation of pooling, however, is that it introduces ambiguity to interpreting findings for high-attrition RCTs or studies in which there is baseline inequivalence between treatment and control groups: some audiences will interpret the presence of high attrition or baseline inequivalence at any time point in

[5] See Wang and Maxwell (2015) for more information on using multilevel models for longitudinal analyses.

[6] See Fitzmaurice et al. (2012), Diggle et al. (2002), and Vickers (2003) for discussions of statistical power for longitudinal data analysis.

[7] See the final section on threats to validity, which discusses the *missing at random* assumption.

a longitudinal analysis as potentially compromising the credibility of *all* impacts reported in the single model. We unpack this limitation and ways to address it in detail below.

It is important to note that both approaches to estimating impacts (that is, using multiple analytic models or a single longitudinal model) across time points can yield credible impacts. However, a given study's context, assumptions, and operationalization might warrant use of one approach over the other.

Table 2 summarizes things to consider in deciding how to model impacts over time.

Table 2. Decision 2 considerations

Decision	Things to consider
2. How to estimate impacts across time points	Statistical power can be higher in longitudinal models due to the assumption that covariates are constant across all assessment periods.
	If understanding the persistence of impacts across time periods is central, longitudinal models can be a more credible approach because an inferential test of the difference in impacts across time points is readily available.
	Establishing credibility of findings in longitudinal models when there is high attrition and baseline inequivalence in one or more time points could depend on the audience. Different audiences might want to see different analyses to address their concerns, and the simpler analysis of estimating impacts separately for each time point might be more compelling for some.
	Both approaches can yield credible findings.

Decision 3: Treating time as discrete or continuous in longitudinal models

If a study uses a single longitudinal model to estimate outcomes for all time points, a third decision researchers and evaluators will encounter is how to model impacts over time—either as discrete periods that represent point-in-time estimates of program effectiveness, or by treating time as continuous and assessing differences in trends of outcomes.

Modeling time as a discrete variable

When an analysis treats time as a discrete variable, the researcher can calculate an estimate of program effectiveness at each follow-up time without making any explicit assumptions on how those estimates relate to one another. Such a study typically converts time into a series of indicator variables or dummy variables, and the actual amount of calendar time between successive measurements does not play a role in estimation. Models typically require two parameters for each time point: one to represent the expected level of the outcome in the control group at that time, and the other to represent the difference between treatment and control (treatment effect) at that time. Thus, the number of model parameters will increase as the number of measurement times increases. We discuss model specification in more detail in later sections of this brief.

Modeling time as a continuous variable

An alternative approach to estimating impacts over time is to treat time as a continuous variable. The resulting model is known as a growth curve model. In growth curve modeling, the researcher specifies a specific functional form for how outcomes change over time (and how impacts vary with time). The simplest (and most common in the TPP literature) functional form is to assume impacts change linearly over time. Other models could assume a higher-order polynomial relationship (such as quadratic or cubic) or use a nonparametric specification that assumes an unspecified smooth relationship over time. Researchers and evaluators should take care in choosing the most appropriate functional form for impacts over time given theory and data constraints (because higher-order functional forms will be infeasible with a small number of survey assessment points), as this assumption will affect impact estimates.

Growth curve models estimate different outcome trends over time among the control and treatment groups. One type of comparison that researchers might make is to document how the treatment and control groups compare with each other on these trends, and a direct interpretation of model parameters will provide this comparison. Importantly, however, the model parameters from a growth model will *not* directly translate to impacts at specific measurement times (for example, the difference in outcomes at six or 12 months post-program). Researchers can calculate these point-in-time impact estimates through linear combinations of model parameters (and these post hoc point-in-time impact estimates are likely the impact estimates many audiences will look for). This is an important reporting consideration for growth models as evidence reviewers (including the TPPER) will want to see point-in-time estimates of program effectiveness at a well-defined time point (see section on "**Estimating point-in-time effects**" in the appendix for more information).

In contrast with what happens when modeling time as a discrete variable, the number of parameters in a growth curve model does not increase as the number of observation times increases, because the relationship between successive impacts is assumed. Thus, growth curve models can be more parsimonious (require fewer parameters) than discrete-time models when the number of observation times is large. That said, most TPP studies have a relatively small number of follow-up times, and for these studies we recommend treating time as discrete unless there is specific interest in growth trends. When the number of follow-up times is five or more, a growth curve approach is a viable alternative that can result in a more parsimonious model.

Table 3 summarizes considerations for decisions on classifying time.

Table 3. Decision 3 considerations

Decision	Things to consider
3. How to model time (whether to supplement a discrete conceptualization of time with a growth curve)	When conducting growth curve models, estimating a point-in-time difference in outcomes is a best practice for showing the effectiveness of the program, because it goes beyond simply focusing on the differences in trends across the treatment and control groups.
	When there are several assessment periods, growth curves can increase statistical power by consolidating multiple assessment periods in a single trend (however, a study having several assessment periods is unlikely to be the case in many TPP studies).
	In the case of studies with a small number of follow-up assessments (as is the case in most TPP studies) researchers should in most cases treat time as discrete unless there is specific interest in growth trends (which should still be accompanied by point-in-time estimates).

Decision 4: Accounting for within-person correlation in longitudinal models

Traditional regression models often assume that observations are independent. Models for longitudinal data violate this assumption because they include multiple observations for the same individuals. Failing to account for within-person correlation in standard regression models will result in underestimating the uncertainty around impact estimates, which will yield artificially small standard errors (and small *p*-values) for the tests of program effectiveness.

There are three common approaches to addressing within-person correlation in longitudinal models that yield valid standard error estimates, and all three yield results that align with TPPER standards:

- Linear or generalized linear models with cluster-robust standard errors, using the sandwich variance estimator (Huber 1967)
- Generalized least squares (GLS) or generalized estimating equations (GEE)[8]
- Mixed effects models (also known as random effects or hierarchical linear models)

There are several differences between these three approaches. First, we note that the sandwich estimator and GLS/GEE approaches are known as marginal models, whereas the mixed effects models are conditional models. This results in a subtle difference in the interpretation of impacts. Whereas marginal models estimate *the average difference between the treated and control groups*, conditional models estimate *the difference between treated and control for an average person*. This distinction is typically relevant only for nonlinear models, such as logistic or Poisson regression models.

Second, GLS/GEE is the only one of the three listed approaches that allows the estimated within-person correlation to depend on the ordering of the time points; the other approaches assume the measurements are exchangeable (the same correlation between any pair, regardless of when they were measured in

[8] GLS is an extension of ordinary least squares to account for correlated data, and thus applies only when outcomes are assumed to be normally distributed. Nonlinear models (such as logistic or Poisson regression) can also apply GEE, which is somewhat equivalent to GLS in the case of normally distributed outcomes.

time). To do this, GLS and GEE require the user to specify a working correlation structure that is used in estimation. Properly specifying the appropriate working correlation structure can result in lower standard errors (increased statistical power) for the resulting impact estimates, though in practice this effect is usually small compared to the sandwich estimator approach. In addition, we note that the sandwich variance estimator can produce more robust standard errors than from a GLS/GEE model, which is a recommended best practice whenever the analysis cannot confidently assume the true correlation structure.

Finally, we note that generalized estimating equations use an approach known as quasi-likelihood, rather than the full likelihood approach used in generalized linear models and mixed effects models. As we will discuss later in this brief, this means GLS/GEE models might produce biased estimates when there is a large amount of missing data (potentially resulting from attrition over time).

A more thorough discussion of the trade-offs between these approaches is available in Diggle et. al. (2002); Gardiner et al. (2009); and Fitzmaurice et. al. (2012). Regardless of which approach they use; authors should make sure to clearly document it when summarizing their methods and results.

Table 4 summarizes considerations for decisions about accounting for within-person correlation.

Table 4. Decision 4 considerations

Decision	Things to consider
4. Calculating standard errors appropriately, given multiple observations within an individual	There are many options for adjusting standard errors for analyses that include multiple observations within individuals. If the analysis accounts for some adjustment for standard errors across observations nested within individuals, correlated errors will not undermine the credibility of findings.
	A key factor to consider in selecting a strategy that adjusts for multiple observations within individuals is the use of marginal models vs. conditional models, particularly in studies with large amounts of missing data.

Section A of the technical appendix gives technical details on model specification and ways to structure data to align with a given analytic approach.

Threats to validity with repeated measures studies

To produce internally valid impact estimates, rigorous impact evaluations must have treatment and control groups that are equivalent at baseline. Ideally, this is a result of the balance produced through a random assignment process. A common threat to the credibility of study findings is sample attrition, which negatively affects baseline equivalence in the analytic samples. Typically, the study sample will decrease in size over time because participants lose interest in the study or become lost to follow-up. If treatment and control group members leave the sample over time at different rates (a situation known as differential attrition), then the sample that contributes data used in the estimation of program effectiveness (the analytic sample) might not be equivalent on baseline characteristics, even if the full study sample was equivalent at the start of the evaluation. Therefore, researchers must document the extent to which

attrition and baseline inequivalence are a concern for a given study and conduct impact analyses that mitigate these threats, thus ensuring the impact findings they estimate are credible.

Presenting information to satisfy different audiences

Different audiences (including, but not limited to, peer reviewers, evidence reviewers, evaluators, and researchers) could be more or less accepting of the following methods as options to address high attrition or baseline inequivalence in studies with multiple follow-up periods. The following are useful best practices to consider for reporting, regardless of decisions and ultimate approaches for addressing internal validity threats:

1. Present complete information about analytic sample sizes at each time point to enable readers to assess attrition. Some audiences will infer that low attrition at a given time point will maintain the initial equivalence produced through random assignment.
2. Present information on baseline equivalence for the analytic samples at each follow-up period and at the initial baseline period. Some audiences will see equivalence on measurable characteristics of participants as a necessary condition for producing a credible impact estimate at a given assessment period.
3. Conduct benchmark and sensitivity analyses using the different analytic approaches suggested earlier to address threats to internal validity. This will demonstrate the robustness of your findings to different analytic approaches and different ways of addressing potential threats to internal validity.

We provide more guidance on these topics in the subsections below.

Reporting attrition and assessing baseline equivalence

Researchers working on any TPP study should document response rates and baseline equivalence of analytic samples at each follow-up period to communicate the presence of internal validity threats in an impact study (see Cole and Chizeck [2014] for guidance on this topic). Reviewers of the study findings can use this information to examine whether overall and differential attrition are sufficiently low (and thus do not threaten internal validity)—for example, by examining attrition rates relative to the TPPER attrition standards (U.S. Department of Health and Human Services [HHS] 2022).

When attrition is high, documenting nonresponse is not enough to satisfy most audiences. In these cases, researchers should assess baseline equivalence of the analytic samples used to estimate the effects of the program at all follow-up periods. Even in cases with low-to-moderate attrition, demonstrating baseline equivalence for the sample that remains at each time point is good evaluation practice for RCTs, and is necessary for all quasi-experimental analyses. Assessing baseline equivalence is straightforward; see Cole and Agodini (2014) for recommendations on approaches for conducting the analysis, and HHS (2022) for guidance on current standards used to assess the equivalence of groups.

For purposes of this section, *inequivalence* refers to a baseline difference greater than 0.25 standard deviations on a characteristic required by TPPER standards. TPPER examines studies for equivalence on three demographic characteristics: age or grade level, biological sex, and race/ethnicity. In addition, for studies with sample members in eighth grade or higher, the study authors must also establish baseline equivalence on sexual behavioral outcome measures (for example, rates of sexual initiation).

Addressing baseline inequivalence at one or more follow-up times

Producing credible estimates of program effectiveness will require extra work when a study has high attrition and/or baseline inequivalence at any follow-up period. When attrition (or other key design features) results in baseline inequivalence of any potential analytic samples, standard statistical adjustment approaches (such as regression adjustment) applied to either the full observed data (available data analysis) or the subset of respondents with observed data at all time points (complete case analysis) can yield biased impact estimates.

Several available analytic methods can, when used correctly, mitigate or remove this bias. Unfortunately, each of them makes a key assumption about the missing data mechanism: that the missingness of any particular observation does not relate directly to the true, unobserved measurement value and that other variables included in the analysis can predict the missingness instead. This assumption is known as *missing at random*. The opposite scenario (*not missing at random*) occurs when the missingness is directly associated with the value itself—for example, if individuals with worse outcomes at a particular time point are more likely to drop out, independent of past observations or other information in the model. Because this assumption is untestable, researchers should use each of these methods with caution. See Deke (2013) or Kautz and Cole (2017) for guidance on imputation and benchmark and sensitivity analyses to handle missing data in TPP studies as one way to address this concern. Next, we discuss two other, more nuanced considerations in the longitudinal analysis space.

Matching and weighting

There are alternatives to imputation as a means to address and possibly ameliorate baseline inequivalence (potentially stemming from sample attrition). One approach is to use matching or weighting methods to manufacture equivalence of the analytic sample at later follow-up points.

The use of propensity scores is one common approach: researchers use background or baseline variables to estimate individuals' propensity to be members of the treatment group and then either match them based on their propensity scores or weight the data for each observation based on the inverse of the propensity score. There are many other approaches for matching and weighting; see Cole and Agodini (2014) for recommendations on how to conduct this in the TPP area.

Matching and weighting approaches are fairly straightforward when researchers estimate impacts separately at each follow-up period (Decision 2), because they can estimate separate (propensity) scores or weights at each time point. Thus, it is feasible to use different matched or weighted samples for each follow-up period as a way to produce credible impacts.

However, for longitudinal analyses, these types of analyses can be more complicated, as there will be different respondents at each follow-up period. Thus, there will be different sets of potential propensity scores and different sample members contributing to the estimation of program impacts at the different time points One simple option for longitudinal analysis when there is inequivalence at one or more time points is to create a complete case sample: the subset of individuals with complete data at all periods.[9] We could estimate a single propensity model for this complete case sample and conduct a matched or weighted analysis on this sample as a means to manufacture equivalence of a single sample. A more efficient approach is to use all available data and produce weights that vary at each observation time. Fitzmaurice et. al. (2012) provide more details on weighting approaches to account for sample attrition.

Note: We do not expect matching and weighting to yield evidence of program effectiveness as credible as the evidence from a well-implemented RCT with low levels of sample attrition. Matching and weighting can eliminate bias associated with selection on observable characteristics, but they cannot eliminate bias due to unobserved characteristics. For this reason, the TPPER standards allow only studies that use matched or weighted samples that satisfy baseline equivalence requirements to receive a moderate evidence rating.

Maximum likelihood-based estimation

Another approach to address bias induced by attrition and baseline inequivalence is to use maximum likelihood-based estimation methods. These methods will provide unbiased estimates of treatment effects under a missing-at-random assumption.

Maximum-likelihood approaches are commonly used when the analysis uses mixed effects models (notably, those that use random effects to account for non-independence among observations nested within an individual) or the analysis uses a structural equation model (typically by analyzing the covariance matrix of the data contributing to the analysis), and thus, these approaches can yield credible impacts when data are missing at random (Singer and Willet 2003).[10]

On the other hand, if a researcher does not use a likelihood-based approach to modeling longitudinal data (like the GLS and GEE approaches discussed earlier in this brief) they would have to rely on matching and weighting or imputation procedures for manufacturing baseline equivalence. Standard generalized linear models are likelihood-based, but they assume independence between observations that likely does not hold in longitudinal data. These methods are potentially biased when differential attrition results in baseline inequivalence.

[9] Some audiences prefer seeing impacts estimated from a single complete case sample as a means to compare program impact estimates over time, net of sample composition changes (because the study will use a single sample to estimate impacts at all time points).

[10] See Singer and Willet (2003) for details on mixed effect and covariance structure analysis approaches to maximum likelihood estimation.

Summary

This brief has reviewed common decision points that researchers and evaluators confront when they conduct longitudinal RCTs. These decision points are:

- **Decision 1:** Accounting for the baseline measure of the outcome, either with the DID approach by adjusting baseline scores on the left-hand-side of the model equation or with the ANCOVA approach by adjusting baseline scores on the right-hand-side of the model equation
- **Decision 2:** Estimating impacts across time points, either with separate models for each follow-up assessment or a single longitudinal model that estimates all time points simultaneously
- **Decision 3:** Modeling time in the analytic model, either only as a discrete variable or by including a growth curve term for the functional form of the relationship between time and the outcome to assess rate of change
- **Decision 4:** Calculating appropriate standard errors, given multiple observations for each individual

To assess whether the impacts of a TPP program or intervention persist over time, a study must have multiple follow-up assessment periods. There are different ways of conceptualizing estimating impacts depending on the research question; assumptions of the research study; preferences of the intended audience; and features of the observed evaluation—notably, levels of sample attrition and baseline equivalence across follow-up periods. Because audiences have different preferences for how this information is presented, no single approach will be perfect, and each one has strengths and limitations. We therefore recommend fully transparent reporting of potential threats to the study's internal validity stemming from sample attrition and issues with baseline equivalence, and conducting benchmark and sensitivity analyses in accordance with decisions about preparing data as a means to address these threats. This information will help improve the credibility of presentations of the persistence and longevity of program impacts of future TPP evaluations.

Suggested citation

This brief is in the public domain. Permission to reproduce is not necessary. Suggested citation: Milless, K.L., Gellar, J. & Cole, R. "Considerations for TPP Evaluations with Multiple Follow-Up Assessments." Washington, DC: Office of Population Affairs, Office of the Assistant Secretary for Health, U.S. Department of Health and Human Services, 2023.

References

Cole, R., and R. Agodini. "Baseline Inequivalence and Matching." Evaluation technical assistance brief. Office of Adolescent Health, U.S. Department of Health and Human Services, 2014.

Cole, R., and S. Chizeck. "Sample Attrition in Teen Pregnancy Prevention Impact Evaluations." Evaluation technical assistance brief. Office of Adolescent Health, U.S. Department of Health and Human Services, 2014.

Deke, J. "Coping with Missing Data in Randomized Controlled Trials." Evaluation technical assistance brief. Office of Adolescent Health, U.S. Department of Health and Human Services, 2013.

Diggle, P.J., P. Heagerty, K.Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

Fitzmaurice, G.M., N.M. Laird, and J.H. Ware. *Applied Longitudinal Analysis* (vol. 998). John Wiley & Sons, 2012.

Gardiner, J.C., Z. Luo, and L.A. Roman. "Fixed Effects, Random Effects and GEE: What Are the Differences?" *Statistics in Medicine,* vol. 28, no. 2, 2009, pp. 221–239.

Huber, P.J. "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,* vol. I, 1967, pp. 221–233.

Imbens, Guido W., and Jeffrey M. Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature,* vol. 47, no. 1, 2009, pp. 5–86.

Kautz, T., and R. Cole. "Selecting Benchmark and Sensitivity Analyses." Evaluation technical assistance brief. Office of Adolescent Health, U.S. Department of Health and Human Services, 2017.

McKenzie, D. "Beyond Baseline and Follow-Up: The Case For More T in Experiments." *Journal of Development Economics*, vol. 99, no. 2, 2012, pp. 210–221.

Singer, J.D. and J.B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence.* Oxford University Press, 2003.

U.S. Department of Health and Human Services. "Identifying Programs that Impact Teen Pregnancy, Sexually Transmitted Infections, and Associated Sexual Risk Behaviors: Review Protocol (Version 6.0)." 2022.

Vickers, A.J. "How Many Repeated Measures in Repeated Measures Designs? Statistical Issues for Comparative Trials." *BMC Medical Research Methodology,* vol. 3, no. 1, 2003, pp. 1–9.

Wang, L.P., and S.E. Maxwell. "On Disaggregating Between-Person and Within-Person Effects with Longitudinal Data Using Multilevel Models." *Psychological Methods,* vol. 20, no. 1, 2015, pp. 63–83. https://doi.org/10.1037/met0000030.

What Works Clearinghouse. *What Works Clearinghouse: Procedures and Standards Handbook (Version 5.0)*. Institute of Education Sciences, U.S. Department of Education, 2022. https://whatworks.ed.gov.

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