Local Evaluation Technical Assistance

Instructions for Impact Evaluation Analysis Plan Template for HMRE Award Recipients

Month, Day, Year

Mathematica® Inc

This page has been left blank for double-sided copying.

Contents

This page has been left blank for double-sided copying.

The Impact Evaluation Analysis Plan

The Administration for Children and Families (ACF), Office of Family Assistance (OFA) is requiring that all Healthy Marriage and Relationship Education (HMRE) award recipients with local impact evaluations funded by OFA provide an analysis plan for their evaluations.

The objective of an impact evaluation is to test the effect of an intervention, or a component of an intervention, by comparing the outcomes of people who were assigned to be offered the intervention with outcomes of people who were assigned *not* to be offered the intervention.[1,2]

The impact analysis plan is a document that describes the proposed research questions, the selected outcome measures, the program design and counterfactual conditions, the evaluation study design, and the proposed analytic approaches to gauge the effect or impact of the intervention. In addition to helping you document your approach for the impact analysis, this plan helps you detail information that you can use in your final evaluation report or other dissemination products. It builds on the most recently approved evaluation design plan and further refines it.

The reason for developing an analysis plan, before conducting any analysis, is that it demonstrates a team's commitment to being objective with a prespecified, systematic, and scientific approach. It also promotes transparency and credibility by showing your team preselected outcomes and analytic approaches to gauge program effectiveness, thus assuring ACF, program staff, and other interested parties that you have not focused on outcomes that happen to emerge as statistically significant.

This document provides instructions for completing the analysis plan for an impact evaluation. The award recipient's local evaluation team must provide information on all sections. Please use the provided template (**Impact Evaluation Analysis Plan Template for HMRE Award Recipients**) for this analysis plan. In addition, the evaluation team must complete an implementation analysis plan. Instructions and a template for the implementation analysis plan are in accompanying documents. ACF strongly encourages evaluation teams to share this analysis plan with the award recipient's team, and perhaps even with program staff, so that everyone understands the plan and has an opportunity to discuss key decisions. The analysis plan can be considered an agreement between award recipients and their evaluators on two key aspects: (1) the outcomes the evaluation will examine and (2) the approaches the evaluation team will use to assess program effectiveness on those outcomes.

The instructions presented below are organized as follows:

[1] Those not offered the intervention may receive no services or different services.

[2] Assignment is most often random, though other rigorous designs—such as high-quality, quasi-experimental designs—may employ other methods.

- Sections A to C describe the proposed research questions, the intervention and counterfactual conditions, and the study design. Explaining the intervention and evaluation design is critical for ensuring a consistent evaluation approach and for documenting any changes that may have occurred during the implementation of the intervention and evaluation.
- Section D provides the blueprint for the primary and sensitivity analyses that the team will use to examine intervention effectiveness, and for the exploratory analyses used to examine the secondary research questions.

These instructions have been created so that the evaluation teams of all award recipients conducting an impact evaluation can fill out each section of the analysis plan regardless of the specifics of their evaluation. However, some teams may need to adapt some subsections to fit their design.

Under the direction of ACF, your Evaluation Technical Assistance Partner (ETAP) liaison will review your analysis plan to provide input and support as you draft it. Please email your analysis plan to your Federal Program Specialist (FPS) and copy your ETAP liaison by [INSERT DATE]. For consistency, please use this common file-naming convention when submitting your analysis plan:

[**HMRE Award Recipient Name**] HMRE Impact Evaluation Analysis Plan.docx

Your FPS and your ETAP liaison will review the analysis plan, provide comments and suggested edits, and return it to you to revise. Your revised analysis plan must be approved by your FPS.

Instructions for completing the impact evaluation plan template

ACF expects that evaluators will complete the analysis plans, with input from program directors and/or program staff as appropriate. For this reason, these instructions are mainly directed toward evaluators and include a few technical terms. For many of the sections below, evaluators can draw from the most recently approved evaluation plan.

A. Research questions and outcome measures

1. Primary research questions

The primary research questions for HMRE impact studies focus on gauging an intervention's effectiveness in improving healthy relationship and relationship education outcomes. Such outcomes may include the status and quality of the couples' relationships, the quality of co-parenting or parenting, and economic stability and well-being. Outcomes might be measured in a variety of ways, such as surveys, direct assessments, and observations. A best practice is to focus each primary research question on how the intervention affects a specific outcome measure at a specific time point (for example, one year after completing the intervention). This approach will clearly connect the outcome(s) and the time point(s) to the intervention's logic model for the theory of change. An example of this practice is the question, "What is the impact of [intervention] relative to [counterfactual] on the support and affection that couples feel toward each other one year after the end of the intervention?" This approach can be followed for the outcomes and time periods most important to the local evaluation.

Best practices for primacy research question

1. Limit the number of research questions to three to five.
2. Choose a follow-up time point, such as three months post baseline or one year post enrollment, for impact estimates.
3. Clearly state whether the treatment and comparison groups are expected to differ or be equally effective on the selected outcomes.◢

Because the likelihood of a false positive—that is, estimating a statistically significant impact when no effect exists—increases with the number of outcomes studied, another best practice is to limit the number of primary research questions. Therefore, you should limit the number of primary research questions to three to five. By setting priorities across the outcomes and time points that are essential for conducting confirmatory tests of intervention effectiveness that align with the intervention objectives and logic model, you can reduce the number of primary research questions to the most essential ones.

Some evaluations intend to show that there are no differences in outcomes between the

intervention and control groups. For example, your evaluation may be gauging whether two modes of program delivery are equally effective. This type of design implies a test of equivalence rather than a traditional hypothesis test, or test of significance, which examines whether an intervention leads to differences in outcomes. This type of equivalence testing is discussed further in Section D.2 of the analysis plan.

2. Secondary research questions

Some research questions that are not as central to the intervention's goals are still important and of interest to the award recipients, researchers, and other interested parties. These are generally considered to be secondary research questions.

Although not required, secondary research questions help award recipients examine other (non-primary) outcomes the intervention might influence and are considered exploratory analyses. For example, secondary research questions could focus on the following:

- Examinations of the outcomes specified in the primary research questions but at time points different from the one specified in the primary research questions (such as immediately after the end of the intervention)
- Examinations of other outcomes different from those specified in the primary research questions (for example, precursors to the healthy relationships outcomes of primary interest)
- Examinations of the relationships among moderating or mediating variables and outcomes, such as the relationship between dosage or participation and outcomes

3. Primary and secondary outcome measures

Describe the specific outcome measures you will use to answer the primary and secondary research questions. If you construct measures from multiple items or variables, describe the survey items you will use and how you will code them to create the measure. In general, you should focus on outcomes that apply to the entire sample, such as attitudes, skills, knowledge, and behaviors.

- Complete Table 1 (sample text is included in italics), describing all measures you will use to answer the primary research questions assessing the impact of the intervention. Include the time periods you will use to assess impacts for these questions. These outcomes should map to your proposed primary research questions. Whenever applicable and possible, provide the properties of the outcome measures, such as reliability and internal consistency.
- Complete Table 2 (sample text is included in italics) for all measures that you will use to answer secondary research questions used for non-confirmatory tests of the effect of the intervention. Include the time periods you will use to assess impacts for these questions. These outcomes should map to your proposed secondary research questions.

Table 1. Description of outcome measures used to answer impact analysis primary research questions

Research question #	Outcome name	Description of the outcome measure and its properties	Source of the measure	Timing of measure
	Level of affection	*The outcome measure is a scale (value range 1 to 5) calculated from both partners' responses as the average of five survey items measuring support, intimacy, commitment, trust, and friendship.*	*Local follow-up survey*	*Six months after intervention ends*

Table 2. Description of outcome measures used to answer impact analysis secondary research questions

Research question #	Outcome name	Description of the outcome measure and its properties	Source of the measure	Timing of measure
	Relationship skills	*The outcome measure is a scale (value range 1 to 4) calculated as the average of seven items.*	*nFORM exit survey*	*At post-test (immediately after intervention ends)*

B. Description of the focal population and the intervention and counterfactual conditions

1. Focal population(s)

Describe the focal population(s)—that is, provide information on the characteristics of the population that each component of the intervention intends to serve, such as age, gender, marital status, and socioeconomic status. An example of this would be, "The component is intended to be delivered to couples, who are low income, that are in a romantic relationship and have children under the age of 18."

2. Intended intervention condition(s)

Describe the intended experiences of those in the intervention condition(s)—that is, what the intervention aims to offer them. Also, describe any services-as-usual resources available to this group outside of the study. If the program has two or more intervention conditions, provide a separate section for the description of each condition and develop a consistent naming convention (for example, Condition 1, Condition 2, and so on). Describe the following:

- **Intended components.** Describe all the key structural elements of the intervention (for example, group classes, workshops, or one-on-one services) that the members of the intervention groups are meant to receive, and the comparison groups are not. If this is an intervention that includes multiple components, describe all of them. An example would be, "This is a multicomponent intervention in which parenting couples receive classes in

relationship skills, workshops on economic stability topics, case management, and booster sessions." If the intervention consists of adding services to a particular program or offering multiple modes of a program (for example, live streaming), describe the program and all the additional services or modes of delivery that will be provided as part of the intervention. If the intervention consists of providing a number of services not related to a curriculum or program (for example, case management, counseling, or home visits), describe each of the services.

- **Intended content.** Provide the name of the curriculum used (if any) and describe the topics the intervention covers and the resources and materials provided.
- **Planned dosage and implementation schedule.** Describe the number of sessions and the duration of each component of the intervention. Include the length of each session and how frequently the sessions occur. An example would be, "This is an eight-month workshop, with sessions occurring once a week for two hours per session." Describe variation in the frequency or length of sessions across sites, if applicable.
- **Delivery mode.** Describe where the intervention component takes place and who delivers it.
- **Staff characteristics, education, and training.** Provide staff characteristics, such as gender, cultural background, required or typical education level, the hiring requirements of the providers or facilitators of each component, and the training and technical assistance offered to providers before they begin to deliver a component and periodically afterward to maintain fidelity.

Tables can be used to clearly and succinctly summarize intervention components. See Tables 3 and 4 below for an example (sample text is included in italics). If there are multiple intervention conditions, consider using separate tables to summarize each condition.

3. Intended counterfactual condition

Describe the intended experiences of those in the control/comparison group—that is, those in the counterfactual condition. If the award recipient (or partner) is not providing services to people assigned to the control/comparison group, describe any services-as-usual resources available to this group outside of the study. If the control/comparison group is receiving an alternative intervention, describe the following:

- **Intended components:** Group classes, workshops, one-on-one services, and the like
- **Intended content:** Curriculum, topics it will cover, and resources and materials participants will receive
- **Intended dosage:** Total intended dosage, number of sessions and their length, frequency of sessions or services, time period during which services take place
- **Delivery mode:** The setting where the alternative intervention will take place and who delivers it, the intended characteristics of the alternative intervention providers, and the training and technical assistance providers will receive

If the control group received a delayed intervention (also called a wait-list control design), describe any services or interactions with intervention staff that the control group received

during the evaluation period while they were on the wait list. You do not need to describe the delayed intervention this group received, because it occurred after all evaluation-related data collection ended.

Summarize the counterfactual services components in a table. An example is in Tables 3 and 4 below.

Table 3. Description of intended intervention and counterfactual components and focal populations

Component	Curriculum and content	Dosage and schedule	Delivery	Focal population
Intervention				
Relationship skills workshops	*Healthy relationships curriculum: Understanding partner's perspectives; avoiding destructive conflict; communicating effectively*	*Twenty hours, with two-hour sessions occurring twice a week, or four-hour sessions occurring every Saturday*	*Group lessons provided at the intervention's facilities by two trained facilitators in every session*	*Married couples with low incomes*
Economic stability workshops	*Resume preparation; interview and communication skills; appropriate work attire; financial literacy*	*Monthly two-hour workshops*	*Workshops are provided by one facilitator*	*Individual members of the couple who need job search assistance*
Counterfactual				
Economic stability workshops	*Resume preparation; interview and communication skills; appropriate work attire; financial literacy*	*Monthly two-hour workshops*	*Workshops are provided by one facilitator*	*Individual members of the couple who need job search assistance*

Table 4. Staff characteristics, education, training, and development to support intervention and counterfactual components

Component	Staff characteristics, education, and initial training	Ongoing staff training
Intervention		
Relationship skills workshops	*Facilitators are male and female and hold at least a bachelor's degree and received four days of initial training.*	*Facilitators receive a half day of semiannual refresher training in the intervention's curricula from study staff.*
Economic stability workshops	*Facilitators are male and female and hold at least a bachelor's degree and received two days of initial training.*	*Facilitators receive a half day of semiannual refresher training in the intervention's curricula from study staff.*
Counterfactual		
Economic stability workshops	*Facilitators are male and female and hold at least a bachelor's degree and received two days of initial training.*	*Facilitators receive a half day of semiannual refresher training in the intervention's curricula from study staff.*

4. Services actually received by the intervention and control/comparison groups

After describing the intended intervention and control/comparison group services, the analysis plan should provide information on what was actually received by people in the two study groups. Describe plans for measuring the services actually received by each group by following the instructions in the implementation analysis plan document and template. These data will be used to provide context for the impact estimates.

C. Study design

Describe the focal population for this evaluation, how you are recruiting the participants, and the adopted study design (RCT or QED) used to assign participants to study conditions (treatment vs. control).

1. Evaluation enrollment and assignment to study conditions

Describe how members of the focal population become part of the impact study sample. Provide information for the full sample (both intervention and control/comparison groups) on study recruitment, target sample size, eligibility requirements, special recruitment or enrollment procedures (if any), and the consent process. Include information on the following:

- **Recruitment and study sample enrollment targets.** Describe where participants were recruited, including agencies and schools and all service locations or sites. Note any differences in recruitment locations by intervention and control/comparison groups. Provide the target sample size for the study separately by intervention and control/comparison groups.
- **Participant eligibility criteria.** Describe any required characteristics for sample inclusion (for example, age, marital status, involvement with the child support system, attending a particular school, geographical area, and employment status).
- **Special recruitment and enrollment procedures.** Describe any additional criteria for recruiting and selecting the sample beyond the eligibility criteria (for example, if the sample is composed of a random selection of eligible participants, or if the sample includes only specific classrooms in eligible, participating schools).
- **Consent process.** Provide the name of the Institutional Review Board that approved the study design and data collection plans, the date of approval, and the dates of any supplemental review approvals. Describe, in detail, the consent process for both the treatment and control/comparison groups (and, if your study involved underage youth, the process for adult consent and youth assent). Include descriptions of similarities and differences between groups with respect to timing, process, and materials used (such as consent forms or incentives). In general, consent should be obtained before assignment to conditions. However, some cluster RCTs and QEDs may be exceptions if the clusters were assigned to condition before individuals within those clusters provided consent—for example, if classrooms within schools were randomly assigned before the beginning of the school year. If you randomly assigned people before the consent process, note whether

you informed potential sample members of their condition before or after receiving their consent.

Randomized controlled trial and random assignment process

Describe the following about the random assignment process:

- What is the unit of randomization (for example, individual clients, couples, agencies, schools)?
- Who randomly assigns units to the conditions (treatment or control), and when, how, and under what circumstances does this occur?
  - Do evaluation staff or intervention staff conduct the process to randomly assign units to conditions?
  - When does random assignment occur with respect to the timing of consent and baseline data collection? (If using nFORM data, this could include administering the applicant characteristics survey.) For clustered randomized controlled trials (RCTs), who, if anyone, learned of the outcomes of random assignment before consent and baseline data were collected, and for what purposes? Ideally, participants should learn about their randomization status only after they have consented and have completed the baseline survey.
  - What is the method of random assignment (such as random number generation in Excel)?
    - Does randomization occur all at once (that is, the study randomly assigns a large number of units at a single point in time) or on a rolling basis (that is, the study randomly assigns small numbers of units at different points in time)? Describe the details of this process.
  - Describe any stratification or blocking you used to create separate instances of random assignment in the evaluation. For example, you might randomly assign people to a condition separately across service locations, such as schools; in this situation, the service locations are strata or blocks.
- Report the intended probability of assignment to the treatment group. If it varies systematically (for example, across blocks or strata), report why and give the range of probabilities used.
- If applicable, describe any subsampling that occurred after random assignment, the reason for it, the criteria used, and how you implemented the subsampling.

Quasi-experimental design and research group formation

Describe the process you used for identifying and forming the treatment and comparison groups, including whether you assigned clients or groups of clients to the treatment or comparison group. Specify when this assignment procedure occurred, relative to the timing of obtaining consent and collecting baseline data.

If an administrative data set was used to identify the comparison group, describe the source

of the data and the criteria for identifying people similar to the clients in the treatment group, including characteristics and variables used to create comparable groups. Describe any services people in this group have received that are similar to the treatment services.

2. Data collection

Describe the data sources for the analyses. Describe the timing of each data collection point (for example, baseline and the follow-up periods used for primary and secondary research questions). Describe the modes and methods of collecting data at each data collection point (for example, in-person paper survey, online survey, and the party responsible for collecting data at each time point and for each study condition). Thoroughly describe the process and the timing for data collection, by study condition (treatment and control/comparison) and for each data collection time point. Clearly articulate similarities and differences across study conditions and time points. Please use a table to clearly and succinctly summarize features of the data collection procedures for each study group and time point (see Table 5 for an example; sample text is included in italics). Finally, please provide a copy of your data collection instruments in an appendix to your analysis plan (at a minimum, provide the instruments you are using to collect the outcome data that you will use to answer the primary research questions of your study).

Table 5. Key features of the data collection

Study group	Data source	Timing of data collection	Mode of data collection	Party responsible for data collection	Start and end date of data collection
Intervention	*nFORM entrance and exit surveys*	*Enrollment (baseline)* *End of intervention (eight months after enrollment)*	*In-person online survey*	*Program staff*	*September 2021 through January 2025*
	Local evaluation survey	*Three months after the end of the intervention (11 months after enrollment)* *Six months after the end of the intervention (14 months after enrollment)*	*Telephone survey*	*Evaluation staff*	*August 2022 through March 2025*
Comparison	*nFORM entrance survey*	*Enrollment (baseline)*	*In-person online survey*	*Program staff*	*September 2021 through January 2025*
	Local evaluation survey	*Eight-month follow-up 11-month follow-up* *14-month follow-*	*Telephone*	*Evaluation staff*	*August 2022 through March 2025*

Study group	Data source	Timing of data collection	Mode of data collection	Party responsible for data collection	Start and end date of data collection
		up			

3. CONSORT diagram

ACF requires that analysis plans include a CONSORT diagram. A CONSORT diagram is a flow chart that summarizes the number of clients in the study from initial enrollment through the final data collection point, separately for treatment and control/comparison groups. The CONSORT diagram serves two purposes: (1) to assess the estimated sample size and compare it to the target sample size that was the basis of the study's power calculations and (2) to assess the likelihood of high overall and differential attrition rates for the final analytic sample at key follow-up time periods.

The CONSORT diagram you will include in your analysis plan will be an interim version that will include the most current information about your sample. At a future date, you will need to revise the interim CONSORT diagram to create a final version with details on the final sample. Your final evaluation report will include this final version of the CONSORT diagram.

The CONSORT diagram for your evaluation

There are four types of CONSORT diagrams. Select the correct one for your evaluation by assessing the level at which assignment was conducted (cluster versus individual) and whether consent occurred before or after assignment:

1. Cluster-level assignment and consent before assignment (Figure A.1)
2. Cluster-level assignment and consent after assignment (Figure A.2)
3. Individual-level assignment and consent before assignment (Figure A.3)
4. Individual-level assignment and consent after assignment (Figure A4)◢

Appendix A includes CONSORT diagram templates that you can adapt to your evaluation; see the callout box on this page for information on how to select a template. In completing the template, indicate the date through which you enrolled the sample and the date through which you collected and included survey data that are represented in the counts in this diagram. Keep in mind that participants may be at different stages of the study at any point in time due to rolling enrollment or multiple cohorts of implementation, so some participants might not yet have data at all data collection points.

D. Analysis

The analysis plan for evaluating impacts should clearly describe the following: (1) the steps you will take to prepare the data for analysis; (2) the analytic sample (or samples) used in the analyses; and (3) the modeling approach adopted for both the primary and secondary analyses (which

Best practices for handling missing data for primary analysis

- The analytic sample should only include cases with complete baseline and outcome data.
- Imputation of item-level data may be permissible.
- Completely missing outcome measures ◢

are used to answer the primary and secondary research questions, respectively) for the final report. Even though the findings you will report on will be based on the primary analyses, you may want to perform additional analyses, known as sensitivity analyses, under different assumptions from those made for the primary analyses. Sensitivity analyses can help assess how robust your findings are to the assumptions and decisions made to address the primary analyses.

We note here that all the primary analyses must adopt an intent-to-treat (ITT) approach. This means that you will be estimating the impact of being assigned to the treatment rather than the impact of receiving the treatment. Therefore, program dropouts (for example, study participants who do not complete the program they were assigned to, never initiate receiving the program they were assigned to, or decide to cross over to a different study condition) should be surveyed according to your data collection schedule and included in the analysis based on the condition they were originally assigned to.

1. Data preparation

Describe the proposed approach used to clean and prepare the baseline and follow-up data for analysis. Detail the protocols you will be using for data cleaning and handling missing data (see additional guidance on missing data in the section below), including missing data on constructed scales, if the outcome measures you are using are scales or composite measures. Describe your plan for dealing with inconsistent data. This means, for example, describing your plans for identifying and handling responses that are inconsistent with each other or are seemingly inaccurate, across both baseline and outcome (at post-test and follow-up) surveys. If you are administering surveys to both members of a couple separately, describe the strategies you will use to verify that the answers are consistent, such as checking that both members report the same marital status, and what you will do if they are not consistent. Finally, if you are using data from different sources, also include a description of how you will merge or combine all the data.

Missing data. Describe in detail the methods used to assess the level of missing data. There are two main sources of missing data: (1) survey nonresponse and (2) item nonresponse. Survey nonresponse occurs when a participant does not respond to the survey. Survey nonresponse (at follow-up) causes sample attrition (please refer to the section below for more instructions about sample attrition) and leads to missing outcome data. Item nonresponse occurs when a study participant responds to a survey but does not answer one or more items that the participant is eligible to answer.

For the primary analysis, cases with survey nonresponse cannot be included; the analytic sample should focus on cases with complete baseline and outcomes data. However, there is one exception. For studies with low attrition at the couple level, evaluators can choose to use Hierarchical Linear Modeling (HLM) and include all couples with at least one partner responding at the follow-up as the primary analysis. As an alternative to HLM, teams may use couple-level means of the outcome measures (and use the one responding partner in a couple, as available) and evaluate impacts using standard regression-adjusted approaches.

However, if this method of including only couples with one responding partner is used, the team should also conduct and include the analyses with couples who have complete data as a sensitivity analysis and discuss differences in findings (and in the samples, if they are different).

Cases with item nonresponse may be included in the analytic sample if the missingness can be imputed through an acceptable method. ACF accepts limited imputation of outcome data to answer primary research questions, but only for multi-item scales with 20 percent or fewer items missing. Otherwise, the primary analysis needs to use the sample of participants with complete data or limited imputed baseline data. Different imputation approaches can be used in sensitivity analyses or to answer secondary research questions.

Follow these general guidelines for handling missing data to answer primary research questions:

- Use logical imputation (for example, length of a romantic relationship can be imputed 0 if, in a prior question, the respondent reported never being in a romantic relationship).
- Imputation of missing items is permissible if the items are part of a multi-item scale with 20 percent or fewer items missing.
- Acceptable imputation approaches include mean imputation, hot-deck imputation, regression-based imputation.
- Before proceeding with imputation, share your plan with your ETAP.

2. Attrition and analytic sample

Attrition refers to the number of people in the baseline sample for whom follow-up was not completed or who are missing outcome data. For the purposes of this analysis plan, describe the approach you will use to report overall and differential (between study groups) attrition from the initially assigned sample. ACF follows the What Works Clearinghouse (WWC) standards for computing overall and differential attrition and for determining whether or not the evaluation experienced high attrition. ACF recommends the use of the cautious boundary for all evaluations (with the exception of evaluations serving youth in schools during the regular school day, which can use the optimistic boundary) to determine whether the evaluation experienced high attrition. However, attrition can only be assessed using complete data, not imputed data.

The analytic sample is the sample or samples (there might be as many analytic samples as the number of outcomes used to address the primary research questions) you will use to estimate the impacts of the intervention. Please describe how you will define the analytic sample (for each research question, if applicable) and use the CONSORT diagram as a guide in preparing this description. Clearly describe what data you require for a person to be part of the analytic sample and refer the reader to the specific sections of the CONSORT diagram where you present the number of individual clients participating in the study who meet those data requirements. For example, indicate whether the analytic sample for the study will be individual clients with complete baseline and outcome data for all variables of

interest (that is, a complete-case sample), and refer the reader to the sections of the CONSORT diagram where you present the number of individual clients who completed the baseline, completed the immediate post-intervention follow-up (and subsequent follow-ups, as applicable), and have all the required data so they are included in the primary analysis sample. Note that, because of the intent-to-treat framework, you should not exclude from your analytic sample participants who do not complete services. These study participants need to be included in the analytic sample if they have outcome data (that is, if they complete the data collection efforts), even if they do not complete services.

Assessment of baseline equivalence

It is good practice to assess baseline equivalence for all impact evaluations, even RCTs with low attrition. However, quasi-experimental studies and random assignment studies that lose part of the sample at the follow-up time periods during which they assess intervention impacts (that is, experience attrition) must verify that the study groups (treatment and comparison groups) are equivalent at baseline, because having well-matched treatment and comparison groups reduces the risk of bias in the impact estimates when attrition occurs.

For the purposes of this analysis plan, please describe the baseline measures of the analytic sample you will use to examine the equivalence of the study groups (for example, demographic characteristics and baseline measures of the outcomes of interest). We recommend using effect sizes to assess differences at baseline between the study groups rather than *p*-values, particularly for baseline measures of the outcomes.

Condition crossover

Describe how you will quantify and report the amount of crossover that occurs during the intervention. For example, describe how you plan to use enrollment rosters, workshop attendance data, and survey data to assess whether participants assigned to each study group (treatment and comparison) are receiving services or participating in the activities that are meant exclusively for, or that are substantially similar to, those assigned to the other group. In addition, describe the approach to reporting those instances. For example, explain that you plan to report the percentage of participants in the comparison group who reported receiving relationship skills lessons from any organization. As needed, refer to Table 4 and Section B of the template.

3. Analytic approach

In this section, describe the adopted approaches for answering the primary and secondary research questions and sensitivity analyses.

Analyses addressing primary research questions

This section should include a detailed description of the modeling approach used to answer the primary research

Primary analysis best practices

- Use an intent-to-treat framework that includes all participants randomized to condition in the analytic sample.
- Adopt a point-in-time approach to produce impact estimates. For example, use regression to obtain easy-to-interpret coefficients.
- For evaluations that hypothesize no difference on outcomes between study groups, describe your plan to conduct tests of equivalence as part of your analysis.
- Reserve alternative analysis approaches, such as different model specifications or handling of missing data, for the sensitivity analysis.

questions and the associated sensitivity analyses you plan to conduct, as the final report will focus on the results of these analyses. More specifically, the analytical approach needs to align with the hypotheses associated with the research questions. Typically, the hypothesis associated with the primary research question is that outcomes for participants assigned to the treatment group are better, on average, than outcomes for participants in the control group.

However, some HMRE evaluations have hypothesized that, on average, there are no differences in outcomes between the participants of two treatment conditions; this usually applies when the study is examining the impact of two program delivery modes. If any of your primary research questions assume equivalence of effects between the two study conditions, please refer to Appendix B, which contains guidance on conducting a test of equivalence, and describe plans to include equivalence testing as part of your analysis.

With the intent-to-treat approach, all the study participants who were assigned (randomly, if the study is an RCT) to the study groups (treatment and comparison) are part of the impact analysis even if they did not receive the services they were assigned to receive, and you should analyze them in the groups to which you assigned them.

ACF requires that award recipients produce and report point-in-time estimates to determine the program impacts for ease of interpretation, standardization, and comparison across estimates, and that they reserve alternative modeling approaches for the sensitivity analyses. Point-in-time estimates are obtained via a regression model predicting the follow-up outcome by the baseline outcome and the treatment indicator. You might also need to control for additional variables to account for lack of baseline equivalence or improve precision of the estimates, enabling the evaluation to identify smaller program effects as statistically significant. In addition, to determine statistical significance of the study findings, ACF requires a two-tailed test with .05 significance level (for example, “findings are considered statistically significant based on $p < .05$”). If any of your primary research questions assume equivalence of effects between the two study conditions, please refer to Appendix B, which contains guidance on conducting a test of equivalence, and describe plans to include equivalence testing as part of your analysis.

The description of the analytical approach should include the following information:

- **Modeling approach.** Describe the type of regression model you will use to estimate intervention impacts for each research question (linear regression, logistic regression, and so on).
- **Model specification.** Describe the variables included in the model and the parameters of interest. For example, identify the parameter representing the impact estimate. List all potential covariates you plan to include in the analyses in a table (see Table 6 for an example; sample text is in italics) and justify your reason for including them. Generally, models include measures of the outcome at baseline and demographic characteristics as covariates, because doing so may enhance the precision of the impact estimates.

- If you have not determined the covariates yet, describe a plan for determining those you will include in your analyses. Aside from the baseline version of the outcome of interest, specify whether any covariates will differ across the models used to answer the primary research questions. When appropriate, describe the blocking and stratification variables—for example, county, school size, and cohort—that you will incorporate as covariates. For example, if your evaluation is an RCT with low attrition but the baseline equivalence analysis revealed a difference larger than expected for some characteristics, you might want to include those variables in the model.

/ Specify the statistical software package you will use.

/ Describe how the model will adjust for clustering (if applicable).

Table 6. Covariates included in impact analyses

Covariate	Description of the covariate
Age	*Age (in years) as of the baseline data collection*
Baseline marital status	*Marital status (1 = married; 0 = not married) as of the baseline data collection*

4. Sensitivity analyses for primary research questions

Describe any analysis you will conduct to test the robustness of the results or the appropriateness of the analytic model for the observed data, along with underlying assumptions adopted for addressing the primary research questions. Include analyses that change potentially important research decisions or assumptions. One such example might be analyses using procedures to prepare and handle missing or inconsistent data that differ from the procedures you use in the primary analysis approach (such as alternative methods to impute missing baseline data). Another example might be analyses that adjust for alternative sets of covariates. For instance, the main analysis approach might adjust for covariates at the individual level (such as age, race, or education level) and at the cluster level (such as county, school district, or service location size), yet the sensitivity analyses might adjust for covariates at the individual level only.

5. Analyses addressing secondary research questions

Describe the analytic approach you will use to address all secondary research questions, to the extent that it differs from the analytic approach proposed for primary research questions (for example, you might be interested in conducting a dosage analysis). Please follow the guidance for the primary research questions above.

Researchers can explore a broad range of possible associations among outcomes and mediating factors—without increasing the likelihood of false positives among primary research questions—by differentiating the secondary (exploratory) research questions from the primary ones before analysis begins. In reporting findings, do not highlight findings from the secondary research questions, which are exploratory analyses, even when they are statistically significant. Findings from exploratory analyses are not considered impact findings.

Identify all additional research questions that you plan to address using data from this evaluation. These questions might include secondary, non-experimental analyses on mediator variables, dosage and participation, and the relationship between implementation and impacts. Describe the outcome measures and the planned analytic approaches you will use. The following are examples of research questions for exploratory analyses:

- What is the association between receiving the intervention and outcomes considered precursors to the evaluation's primary outcomes (for example, couples' conflict resolution skills or a precursor to relationship satisfaction)?
- What is the association between receiving the intervention and other outcomes not considered primary, intended outcomes of the intervention (such as obtaining a GED or enrolling in college)?

Appendix A:

Instructions for completing CONSORT diagrams

This page has been left blank for double-sided copying.

Instructions for completing CONSORT diagrams

Appendix A provides instructions on how to complete the provided CONSORT diagram templates based on two guiding principles: (1) whether assignment is at the cluster or individual level and (2) whether consent occurs before or after assignment.

CONSORT diagrams that track clusters as the unit of assignment (if applicable). ACF requires specific information for cluster random assignment and quasi-experimental designs (that is, studies that involve assignment [random in RCTs and nonrandom in quasi-experimental designs] of service locations, community-based organizations, groups of clients, and schools). See the following list for the required information (also see the diagram templates, Figure A.1 for studies in which consent to participation happened before assignment to conditions, and Figure A.2 for studies in which consent happened after assignment to condition):

1. Note the date you are completing the CONSORT template; this indicates the time point of the information.
2. Provide a short paragraph describing how clusters are defined (for example, a group of clients or classroom of students who attended enrollment sessions and consented to participate in the evaluation). Use this paragraph also to describe what makes a cluster eligible for the evaluation, the number of clusters screened, the number of clusters determined to be eligible and the counts and reasons for those screened out, and whether and how you prioritized any clusters for inclusion in the study sample.
3. Indicate the total number of clusters assigned (randomly, if the study is an RCT), the number assigned to each condition (that is, treatment and comparison), and the start and end dates of cluster assignment.
4. Indicate the number of clusters still participating in the study (that is, retained), by study condition, at each data collection time point. A participating cluster is one in which at least one person in the cluster completed the data collection effort.
    a. In addition, note any reasons for clusters dropping out and the number of clusters to which each reason applies.
5. In addition to completing a CONSORT diagram for clusters, **please complete a CONSORT diagram for the individual clients** in participating clusters (see instructions in the next section; templates of the diagrams for individual clients are available in Figures A.3 and A.4 in this appendix).
    b. The primary analysis sample you note in the CONSORT diagrams for both clusters and individual clients is the sample you will use to answer the primary research questions in the final report (that is, the analytic sample; see Section D.2 for more information about the analytic sample) after accounting for sample loss due to attrition, missing data, and any techniques used to establish an equivalent sample at the baseline.

Figure A.1. CONSORT diagram for clusters (if applicable), for studies in which consent occurred before assignment

Complete based on pooled sample to date. All cluster-level studies must also complete the diagram for individual clients.

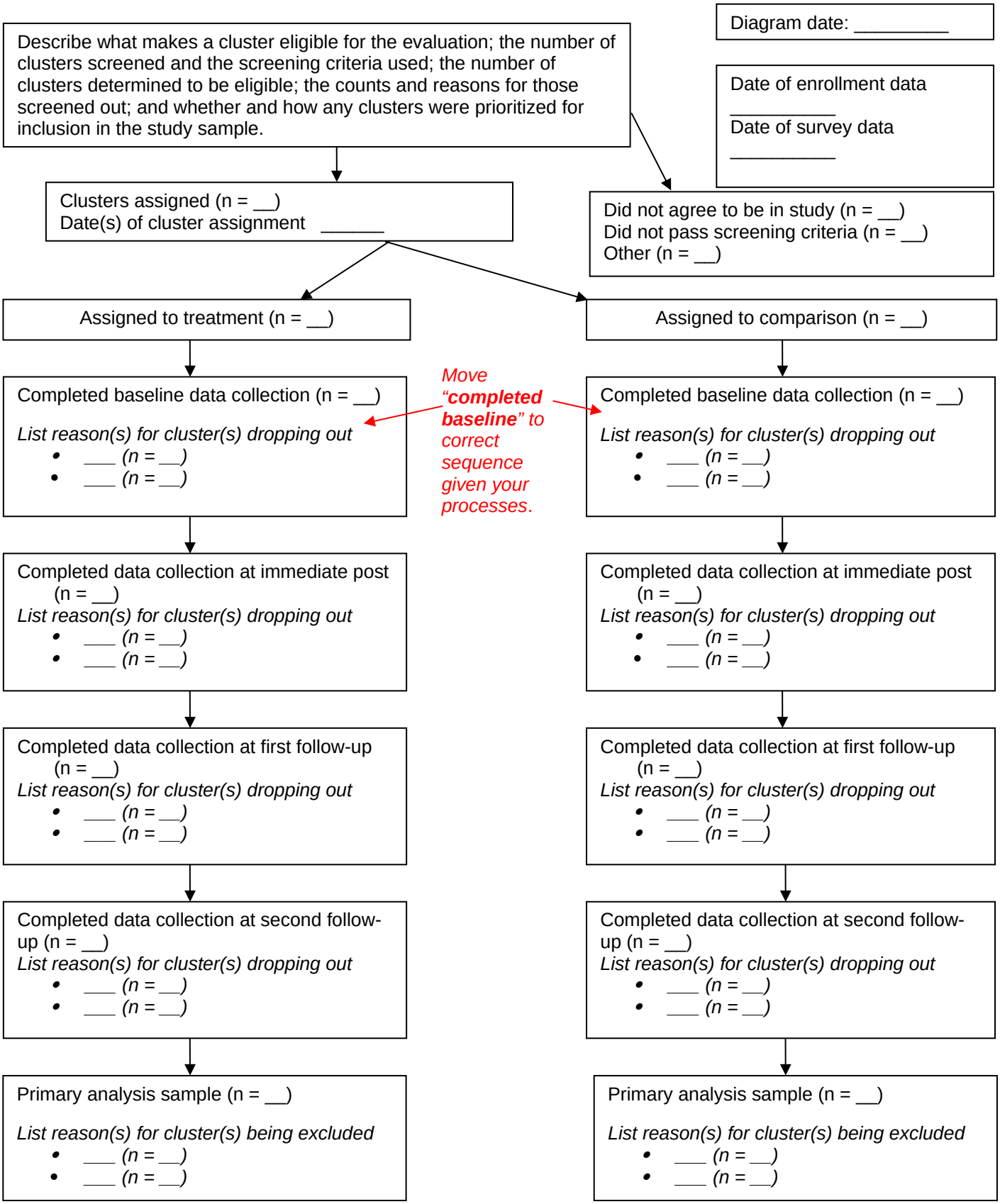
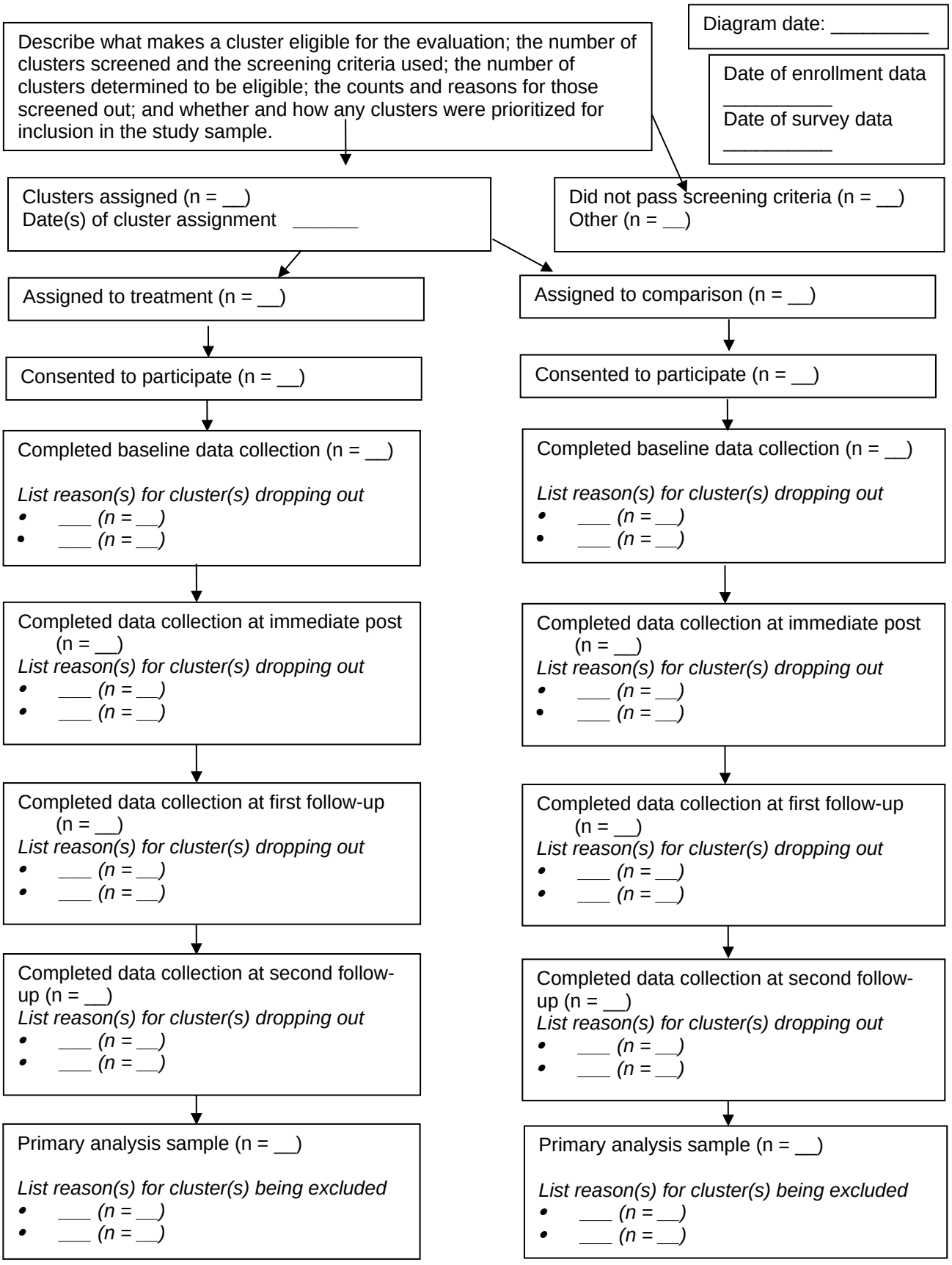


Figure A.2. CONSORT diagram for clusters (if applicable), for studies in which consent occurred after assignment

Complete based on pooled sample to date. All cluster-level studies must also complete the diagram for individual clients.



CONSORT diagrams that track individual clients. Include this diagram for both individual-level and cluster-level designs (see Figure A.3 for studies in which consent to participate occurs before assignment to study condition, and Figure A.4 for studies in which consent to participate occurs after assignment). In addition, provide the following information:

1. Note the date you are completing the CONSORT template; this indicates the time point of the information.
2. Provide a short paragraph describing what makes an individual client (or a couple, for couple-based interventions) eligible for the evaluation. Use this paragraph to also describe the number of individual clients screened and determined to be eligible and the counts and reasons for those who were screened out. In addition, describe the process for selecting study participants among those who were eligible.
3. Indicate the total number of individual clients assigned (randomly, if the study is an RCT), the number assigned to each condition (that is, treatment and comparison), and the start and end dates of assignment.
    a. **If consent to participate in the study occurs BEFORE assignment to condition, please skip to step 6. If consent to participate in the study occurs AFTER assignment to condition, please complete steps 4 and 5.**
4. [If consent to participate in the study occurs AFTER assignment to condition]: Indicate the number of individual clients in the clusters at the time of random assignment, if the study uses a cluster-level assignment.
5. [If consent to participate in the study occurs AFTER assignment to condition]: Indicate the number of individual clients who consented to participate in the study (if you obtained individual consent).
6. Indicate the number of individual clients who provided data, by study condition, at each data collection time point (baseline and subsequent follow-ups).
    a. If the evaluation uses a cluster design, then the number of people in each condition at any time point should reflect the number of people only in participating clusters at that time point. Exclude from these counts people in clusters who have dropped out entirely from the study.
    b. At a given time point, a subset of people may not have been able to contribute data for a particular data collection effort. For example, people who are receiving services and have not yet completed the intervention would not be eligible to contribute follow-up data. Therefore, it is important to document the number of people who are eligible (that is, the number of people who could have contributed data) at a given time point, in addition to the number of people who actually did provide data.
    c. The number of respondents is the number who responded to the survey questions used to measure the primary outcomes specified as your primary research questions. This may be fewer individuals than the number who responded to the survey overall.
    d. Note all reasons for nonresponse and the number of people each reason applies to.

7. Note the intervention start and end dates for the study period.

Figure A.3. CONSORT diagram for individual clients, for studies in which consent occurred before assignment

Complete based on pooled sample to date.

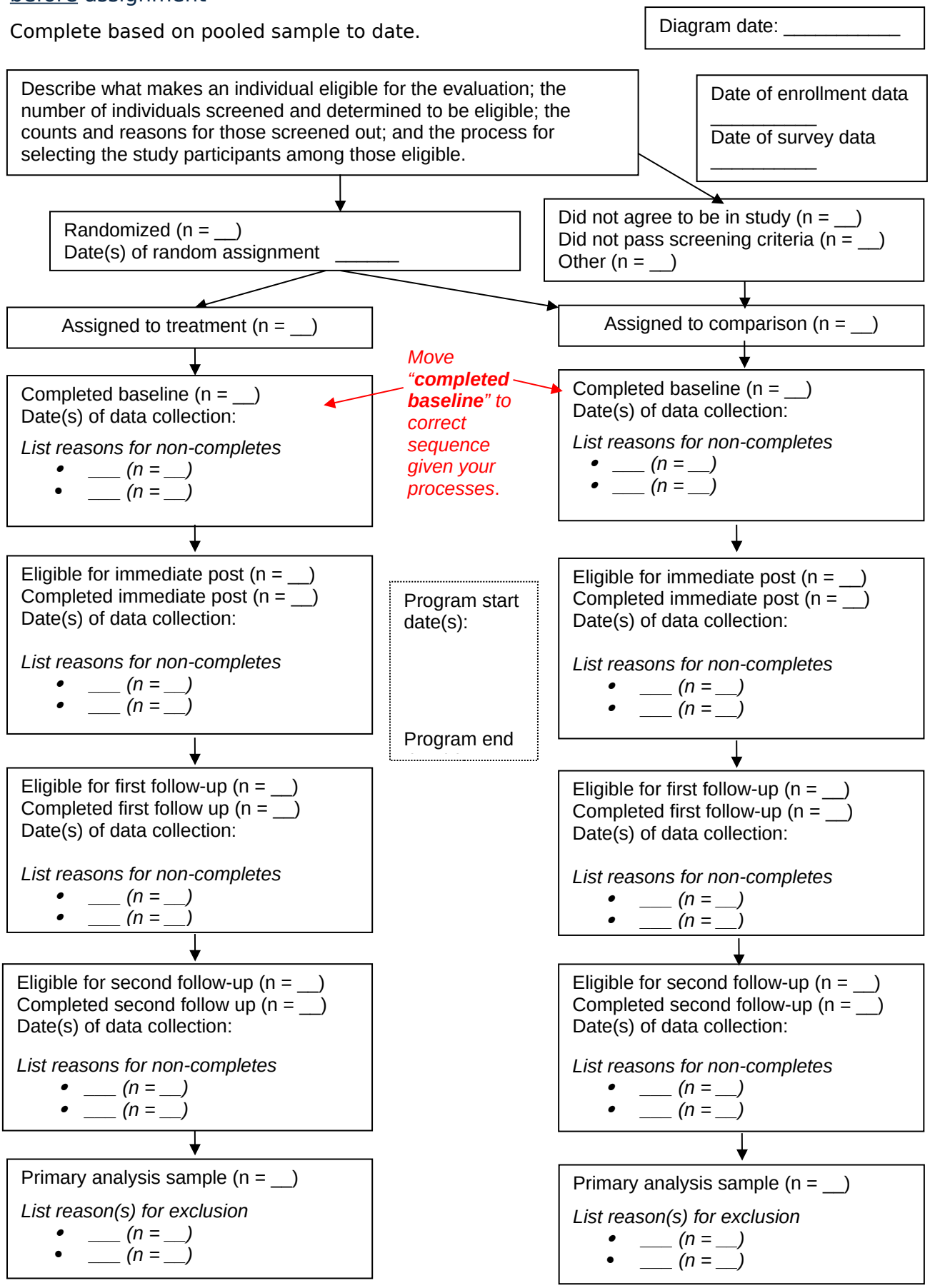
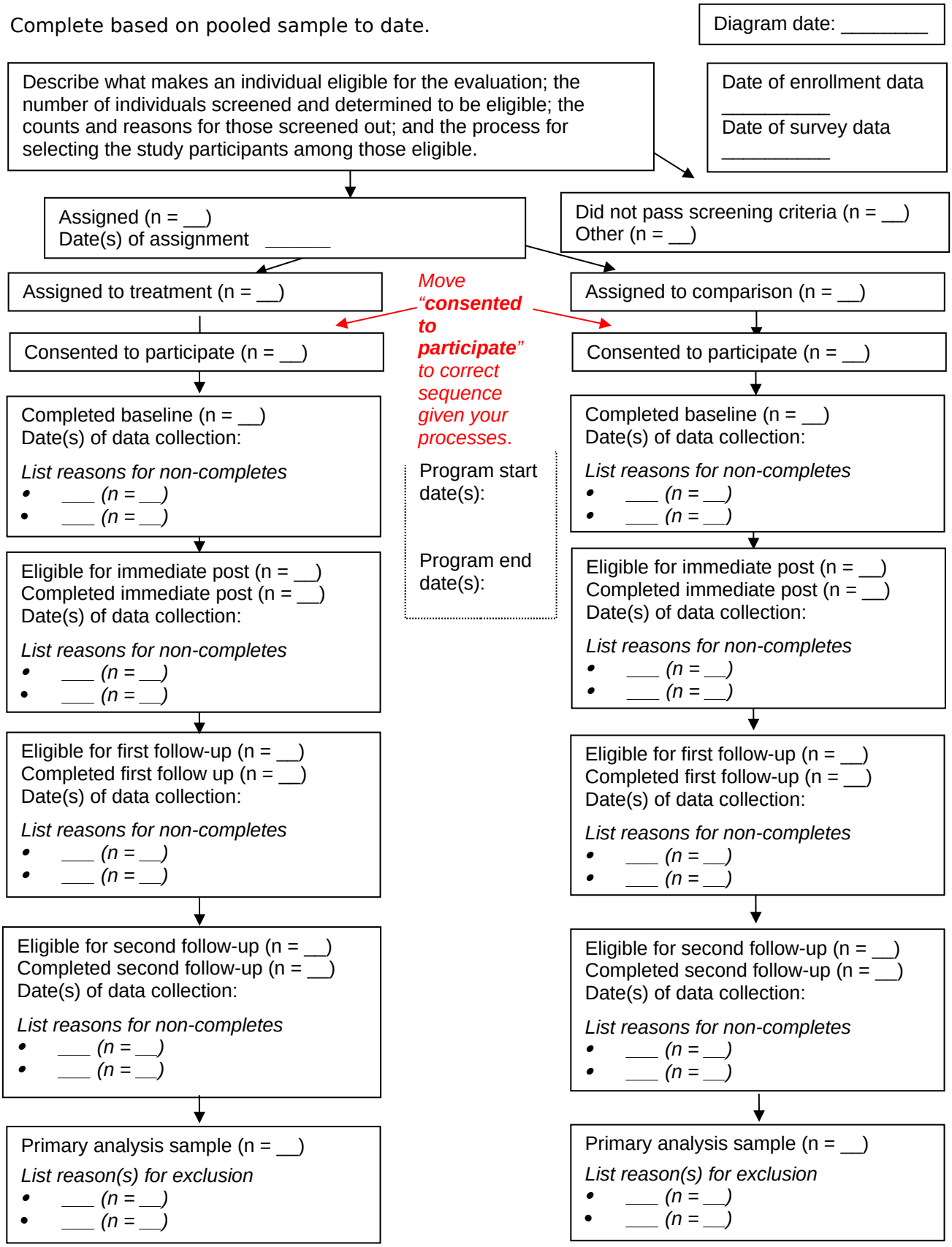


Figure A.4. CONSORT diagram for individual clients, for studies in which consent occurred after assignment

Complete based on pooled sample to date.



Appendix B:

Conducting equivalent effects testing in HMRE impact evaluations

This page has been left blank for double-sided copying.

Conducting equivalent effects testing in HMRE impact evaluations

Section D.3 of this instruction document includes the general guidance for describing the analytic approach in your analysis plan. The instructions note that for studies in which the primary research questions pertain to equivalent effects between the two study conditions, analysis plans should describe how the analysis will test for equivalent effects. The purpose of this appendix is to (1) provide information about testing for equivalent effects and explain how it differs from traditional significance testing, (2) show how to specify the null hypothesis in equivalent effects testing, and (3) demonstrate how to conduct an equivalent effects test.

Your team should use the information in this appendix to inform your analytic plan. Specifically, if your hypotheses assume there will be no differences between study groups, propose an approach for conducting equivalent effects testing in Section D.3 of your written plan. As with the rest of the analysis plan instructions, the guidance in this appendix is mainly directed toward evaluators and includes several technical terms. Box B.1 defines key terms.

Box B.1. Key terms

- **Parameter of interest**: The value that you want to estimate. Common parameters of interest are the mean of a population or the difference in means of two populations.
- **Estimated effect**: The difference in the outcome means of the treatment and control groups. When fitting a regression model, the estimated effect is the coefficient of the treatment group indicator.
- **Null hypothesis (H_0)**: The statement being tested. Typically, in impact evaluations when conducting significance testing, the null hypothesis implies "no difference" or "no effect," meaning there is no difference (or the difference is zero) between the outcome means of the intervention and the comparison groups.
- **Alternative hypothesis (H_a)**: The statement implied if the null hypothesis is not true, that is, that there is a difference in the outcome means between the intervention and comparison groups.
- ***p*-value**: The probability of observing a test statistic (or difference in outcome means) as extreme or more extreme than the value observed when assuming the null hypothesis, H_0, is true. The smaller the *p*-value, the stronger the evidence against the null hypothesis H_0.
- **Two-sample t-test** (also known as the independent samples t-test): A statistical method used to compare the means of two populations.
- **Two-sided hypothesis test** (also known as two-tailed test): A statistical test in which the alternative hypothesis, H_a, states that the parameter of interest is *different* from the value specified in the null hypothesis, H_0. This means that the parameter can be either less than that value or greater than the value specified in the null hypothesis, H_0, but the test does not specify which direction.
- **One-sided hypothesis test** (also known as one-tailed test): A statistical test in which the alternative hypothesis, H_a, specifies that the parameter of interest is greater (or less) than the value specified in the null hypothesis, H_0.
- **Confidence interval** (with a specified confidence level, such as 90 percent or 95 percent): This is a range of estimates for a parameter of interest derived using sample data and a method that has probability equal to the specified confidence level of producing an interval containing the true value of the parameter being estimated. This means that if you were to draw a large number of samples from the same population of interest and construct a 95 percent confidence interval for each sample, 95 percent of the intervals will contain the true value of the parameter of interest. ◢

A. Background

Most Healthy Marriage and Relationship Education (HMRE) impact studies hypothesize that their intervention's estimated impact is statistically significantly greater (or smaller) than the impact of a comparison condition. The standard statistical approach in this case is to test whether the estimated effect, or the difference in the means of the outcomes between the intervention and control groups, is not zero and has an associated *p*-value[3] below a critical threshold, such as $p < 0.05$. However, this approach is *not* appropriate when the key hypothesis is that the impact of the intervention is equal to that of the comparison condition. For example, some HMRE impact studies hypothesize that the same curriculum delivered to both the intervention and comparison groups using two different delivery modalities, such as

[3] The *p*-value is the probability, assuming the null hypothesis is true, of observing a test statistic (or difference in outcome means) as extreme or more extreme than the value observed. The smaller the *p*-value, the stronger the evidence against the null hypothesis.

in-person versus virtually, is equally effective. In this case, using the standard approach of testing for the difference in means between groups may result in a *p*-value that is not statistically significant, such as $p > 0.05$, which may lead you to conclude that there are no differences between the study groups. However, for hypotheses assuming equivalent effects between groups, showing that the differences in means are not statistically significant is insufficient to conclude that two groups produce equal impacts or that the difference in effects is zero. Instead, when testing for equivalent effects, you should propose an alternative approach designed around demonstrating that the differences between two conditions are not large enough to be considered substantively important.

B Hypothesis testing concepts: Comparing significance and equivalent effects testing

To simplify the presentation, this guidance focuses on the scenario in which a two-sided hypothesis test is used to compare the means of two independent samples (an intervention and a comparison group).

When you conduct hypothesis testing, the statement being tested is the null hypothesis. Usually, in significance testing for impact evaluations, the null hypothesis implies "no difference" or "no effect," meaning there is no difference (or the difference is zero) between the outcome means of the intervention and the comparison groups.

The alternative hypothesis (H_a) in significance testing is the statement implied when the null hypothesis is not true—that is, the assumption that the intervention is effective and that there is a difference in the outcome means between the intervention and comparison groups.

By contrast, in equivalent effects testing, the null hypothesis states that the difference in the outcome means is large enough to be of practical importance; this is commonly referred to as the *smallest effect size of interest* (see Lakens 2017). Thus, the alternative hypothesis in this type of testing states that the difference in means of your groups is so small that it is not of practical importance or is less than the *smallest effect size of interest*. The alternative hypothesis is often expressed as an interval (Δ_L, Δ_U), where Δ_L is the lower bound and Δ_U is the upper bound of the difference of the outcome means between study groups. This is called an **equivalence interval**. In most instances, the equivalence interval is symmetric around zero, such as (-0.1, 0.1), but not always. Given the definition of the alternative hypothesis, this means that the null hypothesis implies that the difference in outcome means is outside the equivalence interval, so it is either less than or equal to Δ_L or is greater than or equal to Δ_U.

When you conduct a hypothesis test, regardless of significance or equivalent effects testing, there are two possible outcomes:

1. Reject the null hypothesis and conclude that the alternative hypothesis is true with a prespecified confidence level; or

2. Fail to reject the null hypothesis and state that there is not enough evidence to reject it. **Note that when we fail to reject the null hypothesis, we do not say that it is true.**

Relatedly, two possible errors are associated with hypothesis testing:

1. Type I error: We reject the null hypothesis when it is true.
2. Type II error: We fail to reject the null hypothesis when the alternative hypothesis is true.

Hypothesis tests commonly set the probability of committing a Type I error (statistical significance level) to a small value such as 5 percent and minimize the probability of committing a Type II error. However, the latter depends on several factors such as sample size and the actual difference in means.

When conducting a two-sample t-test (as the examples illustrated thus far show), most statistical software will produce a *p*-value, which is used to determine whether there is enough evidence to reject the null hypothesis. If the *p*-value is smaller than the significance level (or probability of Type I error), you can reject the null hypothesis and conclude that the alternative is true. However, if the *p*-value is greater than the significance level threshold, then you fail to reject the null hypothesis but cannot conclude that it is true. You can only say that there is not enough evidence to reject the null hypothesis. For this reason, traditional significance testing cannot be used when you are testing whether the difference in the outcome means of two study groups is zero. In other words, failure to reject the null hypothesis does not mean that the true difference is zero. Table B.1 summarizes the key differences in the null and alternative hypotheses under significance and equivalent effects testing.

Table B.1. Key differences between significance and equivalent effects testing

	Significance testing	Equivalent effects testing
Null hypothesis H_0	Difference in means is equal to zero	Difference in means is **outside** the equivalence interval
Alternative hypothesis H_a	Difference in means is different from zero	Difference in means is **inside** or equal to the equivalence interval

C. Specifying the null hypothesis in equivalent effects testing

There are different ways for researchers to determine the lower bound (Δ_L) and upper bound (Δ_U) of the equivalence interval to use in their analysis. In this document, we discuss the recommended approach for HMRE impact evaluations. The bounds can be expressed in terms of either effect sizes or differences in means. A best practice is to express the equivalence interval's lower and upper bounds as effect sizes because they are common metrics that are easily interpreted, and then specify the value of the bounds as the smallest effect sizes considered of practical importance for a given outcome measure. The equivalence interval (namely, the alternative hypothesis for equivalent effects testing) should be specified before conducting the analysis.

To identify the *smallest effect size of interest*, you can review previous HMRE impact studies to guide your estimates. For example, you can review the past impact studies from the 2015 HMRE award recipients to get a sense of the sizes of the impacts those studies found. You can also review past federal studies, such as The Supporting Healthy Marriage Evaluation, The Building Strong Families Project, the Parents and Children Together evaluation, and The Strengthening Relationship Education and Marriage Services (STREAMS) Evaluation.[4]

A review of these findings shows that the estimated impacts expressed as effect sizes tend to be small, particularly for behavioral outcomes measured at a six- to 12-month follow-up. In general, effect sizes of 0.2 standard deviations (SDs) are considered large for many HMRE outcomes, and effect sizes as small as 0.1 SDs are often regarded as substantively important.

Thus, when specifying the bounds for the equivalence interval, you should think about the smallest impact that would be considered substantively important for your study. For example, assume that you are evaluating whether curriculum delivery in two modes, live-streaming and in-person, yields equal impacts on relationship commitment outcomes. The outcome measure of commitment is measured on a scale ranging from 1 to 10. You have reviewed prior studies and determined that an increase of one-quarter of a point (0.25 scale points) on the scale score (which is the same as an effect size of 0.10 SD) would be the smallest impact considered substantively important. Therefore, the equivalence interval would be a range, (–0.25 to 0.25) (if expressed in the same measurement unit as the outcome), or –0.10 SD to 0.10 SD (if using the recommended metric of effect sizes). However, note that concluding equivalent effects for a narrow equivalence interval, like in

[4] For STREAMS, see stand-alone impact study reports by Goesling et al. 2022 and Wu et al. (2022).

this example, requires large sample sizes.

D. Conducting an equivalent effects test

To conduct an equivalent effects test, follow the steps below.

- Step 1: Determine the *smallest effect size of interest* by reviewing studies similar to yours.
- Step 2: Specify the equivalence interval.
- Step 3: Conduct the equivalent effects test by doing one of the following:
  - Conduct two one-sided significance tests (known as the TOST procedure); or
  - Estimate the 90 percent confidence interval and compare it to the equivalence interval.

Conducting an equivalent effects test is the same as conducting two one-sided significance tests (with 5 percent significant level), for which the null and alternative hypotheses are defined as follows, respectively:

1. Null hypothesis: difference in means is less than or equal to Δ_L; and alternative hypothesis: difference is greater than Δ_L
2. Null hypothesis: difference in means is greater than or equal to Δ_U; and alternative hypothesis: difference is less than Δ_U

To conclude equivalent effects within the specified lower and upper bounds, both one-sided tests need to reject the null hypothesis. If you cannot reject both null hypotheses, then equivalent effects cannot be established.

Furthermore, concluding that the outcome effects for the intervention and comparison groups are statistically equivalent is like saying that the 90 percent confidence interval for the difference in the means between groups is contained within the equivalence interval bounds. If the statistical software you use does not have a built-in command for conducting an equivalent effects test, you can alternatively compute the 90 percent confidence interval for the difference in the outcome means between groups and compare it against the equivalence interval you specified.[5]

Applying equivalent effects testing

Continuing with the example above, assume that for an impact study testing the equivalent effect of two modes of curriculum delivery, researchers set the equivalence interval to (–0.25, 0.25). Next, the researchers report that the means for the relationship commitment outcome for couples enrolled in the HMRE program are as shown in Table B.2.

Table B.2. Example means and standard deviations of relationship commitment outcome

Curriculum delivery	Mean [a]	Standard deviation (SD)	Sample size (n)
In-person (T1)	9.5	2.0	95
Live-streaming (T2)	9.4	2.1	100

[5] See statistical software resources for equivalent effects testing at the end of this appendix.

[a] Relationship commitment is measured on scale from 1 to 10.

First, the researchers calculate the estimated difference between the two group means, which is 0.1 (mean of T1 – mean of T2), which corresponds to an effect size of 0.05 SDs.

Next, they conduct two one-sided significance tests with the following null and alternative hypotheses:

1. Null hypothesis: difference is less than or equal to –0.25 (or the standardized difference is less than or equal to –0.10 SD); and alternative hypothesis: difference is greater than –0.25 (or –0.10 SD)
2. Null hypothesis: difference is greater than or equal to 0.25 (or 0.10 SD); and alternative hypothesis: difference is less than 0.25 (or 0.10 SD)

Using the TOST procedure, neither test rejects the null hypothesis. The first t-test examining whether the difference is less than –0.25 yields a *p*-value of 0.117. The second t-test examining whether the difference is greater than 0.25 has a *p*-value of 0.305. Because both tests fail to reject the null hypothesis, the researchers cannot conclude that the two group means are equivalent. That is, it is possible the in-person mode and the live-streaming mode have different impacts on relationship commitment outcomes.

Alternatively, if the researchers calculate the 90 percent confidence interval for a two-tailed test, they can compare it to the equivalence interval and reach the same conclusion. In this example, the 90 percent confidence interval, using the same units as the outcome measure, is (–0.385, 0.5853), which is not fully contained in the equivalence interval of (–0.25, 0.25), as both lower and upper bounds of the confidence interval fall outside the bounds of the equivalence interval; therefore, the researchers cannot conclude equivalent effects. Analogously, the 90 percent confidence interval expressed in effect size units is (–0.186, 0.283), and falls outside the range of the equivalence interval of (–0.1, 0.1), meaning the researchers cannot reject the null hypothesis.

It is also noteworthy that, in this example, if the researchers had decided to do a traditional significance test instead of an equivalent effects test, using the standard null hypothesis of "no difference," they would not have rejected the null hypothesis of no difference, as the test is not statistically significant (*p*-value = 0.734). This may have led the researchers to incorrectly conclude that, because the difference was not statistically significant, the groups were equivalent. Instead, the significance test in this scenario only leads the researchers to conclude that there is not enough evidence to reject the null hypothesis—that is, there is not enough evidence to conclude that the estimated difference is statistically significantly different from zero. Failing to reject the null hypothesis does not mean that it is true. Thus, this example further demonstrates the need to use equivalent effects testing in place of significance testing for studies hypothesizing equivalent effects between groups.

Concluding equivalent effects with the TOST procedure requires that both one-sided tests are statistically significant (or, alternatively, that the 90 percent confidence interval for a two-tailed test needs to be fully contained in the equivalence interval). However, it is important to note that concluding equivalent effects typically requires samples sizes that are

much larger than the sample sizes used in the example above. For more information on sample sizes and power considerations when conducting equivalent effects tests, see Lakens (2017).

References

Lakens, D. "Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses." *Social Psychological and Personality Science*, vol. 8, no. 4, 2017, pp. 355–362. https://doi.org/10.1177/1948550617697177

Software resources

Resources for R

https://aaroncaldwell.us/TOSTERpkg/

Resources for STATA

https://www.alexisdinno.com/stata/tost.html

Resources for SAS

https://support.sas.com/resources/papers/proceedings15/SAS1911-2015.pdf

https://support.sas.com/resources/papers/proceedings16/11683-2016.pdf