ASSESSING ATTRITION BIAS

A. INTRODUCTION

In a randomized controlled trial (RCT), researchers use random assignment to form two (or more) groups of study participants that are the basis for estimating intervention effects. Carried out correctly, the groups formed by random assignment have similar observable and unobservable characteristics, allowing any differences in outcomes between the two groups to be attributed to the intervention alone, within a known degree of statistical precision.

Though randomization (done correctly) results in statistically similar groups at baseline, the two groups also need to be equivalent at follow-up, which introduces the issue of attrition. Attrition occurs when an outcome is not measured for all participants who were initially assigned to the groups. Throughout this paper, attrition (missing outcome data) is defined to be the opposite of response (nonmissing outcome data). Attrition can occur for the overall sample, and it can differ between the groups; both aspects can affect the statistical equivalence of the groups. Both overall and differential attrition create potential for bias when the characteristics of sample members who respond in one group differ systematically from those of the members who respond in the other.

To support its efforts to assess design validity, the What Works Clearinghouse (WWC) needs a standard by which it can assess the likelihood that findings of RCTs may be biased due to attrition. This paper develops the basis for the RCT attrition standard. It uses a statistical model to assess the extent of bias for different rates of overall and differential attrition under various assumptions regarding the extent to which respondent outcomes are correlated with the propensity to respond. The validity of these assumptions is explored using data from three past experimental evaluations.

The need for a statistical model of attrition bias to inform WWC standards stems from the fact that other existing attrition standards are not suited to the WWC objective of assessing the validity of RCTs. A number of federal agencies have standards for response rates in data collection, but these standards are intended to be general guidelines for all data collection efforts and, thus, are not tailored specifically to RCTs. For example, the Office of Management and Budget (OMB) and the National Center for Education Statistics (NCES) have established response rate targets of 80% and 85%, respectively. However, these standards do not reference the potential effect of attrition bias within a study on the effectiveness of an intervention. Thus, there are no theoretical or empirical reasons that these thresholds are appropriate in assessing attrition within the framework of WWC standards for study designs.

Prior to the development of the model presented in this paper, the WWC guideline that was in place recognized the need for limits on both overall and differential attrition. Specifically, the standard consisted of fixed limits for the allowable overall attrition (20%) and the allowable differential attrition (7%).[1] However, this standard lacked theoretical and empirical justification. Moreover, it did not recognize any possibility for a tradeoff between overall and differential attrition, such that a higher rate of overall attrition could be offset by a lower rate of differential attrition (and vice versa). These gaps underscored the need for a statistical model on which a standard could be based.

In the next section, we present a framework in which both overall and differential attrition contribute to possible bias. Under various assumptions about tolerances for potential bias, the approach yields a set of attrition rates that falls within the tolerance and a set that falls outside it. Because different topic areas may have factors generating attrition that lead to more or less potential for bias, the approach allows for refinement within a review protocol that expands or

[1] The principal investigator of each review could use discretion in setting an attrition standard for his or her topic area.

contracts the set of rates that yield tolerable bias. This approach is the basis on which WWC attrition standards are set.

B. ATTRITION AND BIAS

Both overall and differential attrition may bias the estimated effect of an intervention.[2] However, the sources of attrition and their relation to outcomes can rarely be observed or known with confidence, which limits the extent to which attrition bias can be quantified. The approach here is to develop a model of attrition bias that yields potential bias under assumptions about the correlation between response and outcomes. This section describes the model and its key parameters. It goes on to identify values of the parameters that are consistent with the WWC's previous standards and assesses whether those parameter values are generally consistent with data from three randomized trials.

1. Model of Attrition Bias

Attrition that arises completely at random reduces sample sizes but does not create bias. However, researchers rarely know whether attrition is random and not related to outcomes. When attrition *is* related to outcomes, different rates of attrition between the treatment and control groups can lead to biased impact estimates. Furthermore, if the relationship between attrition and outcomes differs between the treatment and control groups, then attrition can lead to bias even if the attrition rate is the same in both groups. The focus here is to specify a model showing how bias depends on the correlation between outcomes and attrition and the combination of overall and differential attrition in an RCT.

[2] Throughout this paper, the word *bias* refers to a deviation from the true impact *for the analysis sample*. An alternative definition of bias could also include deviation from the true impact for a larger population. We focus on the narrower goal of achieving causal validity for the analysis sample because nearly all studies reviewed by the WWC involve purposeful samples of students and schools.

To set up the model, consider a variable representing an individual's latent (unobserved) propensity to respond, z. Assume z is normally distributed with mean zero and standard deviation one. If the proportion of individuals who respond is P, an individual is a respondent if his or her value of z exceeds a threshold, z^*:

$$(1) \quad z > \Phi^{-1}\left(1-P\right) \equiv z^*$$

where Φ is the standard normal cumulative distribution function. For example, in a scenario in which 75% of individuals respond ($P = 0.75$), an individual is a respondent if his or her value of z exceeds the value corresponding to the 25th percentile in the z distribution [that is, exceeds $\Phi^{-1}\left(1-0.75\right)$].

The outcome at follow-up, y, is the key variable of interest. We assume that y has a normal distribution. Moreover, we assume that y has mean zero and standard deviation one, given that any variable can be standardized in this way. The relationship between y and z can then be modeled as

$$(2) \quad y = \alpha * z + u$$

where α is the correlation between z and y, and u is a random variable that is independent of z.[3] Note that there are no covariates, and the model assumes no effect of the treatment on the outcome. If α is 1 or –1, the entire outcome is explained by the propensity to respond. If α is zero, none of the outcome is explained by the propensity to respond, which is the case when attrition is completely random.

The correlation between the propensity to respond and outcomes may differ by treatment status. Therefore, we specify Equation (2) separately for treatment and control group members (subscripted by t and c, respectively):

[3] In order for y to be a N(0,1) variable, u must be normally distributed with mean zero and standard deviation $\sqrt{1-\alpha^2}$.

$$(3) \quad \begin{aligned} y_t &= \alpha_t * z_t + u_t \\ y_c &= \alpha_c * z_c + u_c . \end{aligned}$$

Because there is no true impact of the intervention in this model, an unbiased estimator of the impact should, in expectation, find no difference in outcomes between the treatment and control groups. Therefore, in the presence of attrition, *bias* is equal to the difference between the expected values of y_t and y_c among respondents. Based on the properties of truncated normal distributions, the analytic formula for the bias, *B*, is

$$(4) \quad \begin{aligned} B &= E(y_t \mid z_t > z_t^*) - E(y_c \mid z_c > z_c^*) \\ &= \alpha_t E(z_t \mid z_t > z_t^*) - \alpha_c E(z_c \mid z_c > z_c^*) \\ &= \frac{\alpha_t \times \phi(\Phi^{-1}(1-P_t))}{P_t} - \frac{\alpha_c \times \phi(\Phi^{-1}(1-P_c))}{P_c} \end{aligned}$$

where ϕ is the standard normal density function.

Equation (4) shows that bias can be generated by treatment-control differences in the response rates (P_t and P_c) or in the correlation between *y* and *z* (α_t and α_c). If neither *P* nor α differs between these groups, then there is no bias because the same kinds of individuals respond from both groups.[4] However, if response rates differ between the treatment and control groups ($P_t \neq P_c$), then bias occurs even when $\alpha_t = \alpha_c$, because respondents in the treatment and control groups have different average values of *z* and, thus, different average values of *y*. Moreover, if $\alpha_t \neq \alpha_c$, then impact estimates will be biased even if the response rate is the same in both groups; respondents from the two groups will have different average values of *y* stemming from the differences in α .[5]

[4] Those who attrite, nonetheless, will differ systematically from those who do not attrite, which may compromise the external validity of the study. However, we do not address that issue here.

[5] It is possible that a difference in the rate of attrition between groups could offset a difference between α_t and α_c . However, throughout this paper, we conservatively assume the opposite—that these differences are reinforcing, not offsetting.

2. Numeric Relationships Between Bias and Response Rates

From Equation (4), we can map out the numeric relationship between bias (B) and response rates (P_t and P_c) after setting assumptions for the correlations (α_t and α_c) between the propensity to respond and outcomes. Table 1 shows the relationship between bias and response rates for each of several different assumptions about α_t and α_c. Each row of the table pertains to a different combination of response rates in the treatment (P_t) and control (P_c) groups, shown in the first two entries of each row. The remaining entries in each row show magnitudes of bias (in effect size units) resulting from these response rates, with different columns pertaining to different assumptions about α_t and α_c. Therefore, each column of the table maps out a distinct relationship between bias and response rates based on the assumed values of α_t and α_c in the column header.

TABLE 1

BIAS BY RESPONSE RATE AND CORRELATION BETWEEN OUTCOMES AND THE PROPENSITY TO RESPOND (EFFECT SIZE UNITS)

		Level of Bias (effect size units)						
		(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
P_t	P_c	α_t **= 0.27** α_c **= 0.22**	α_t **= 0.32** α_c **= 0.22**	α_t **= 0.39** α_c **= 0.22**	α_t **= 0.45** α_c **= 0.39**	α_t **= 0.55** α_c **= 0.45**	α_t **= 0.71** α_c **= 0.45**	α_t **= 1.00** α_c **= 1.00**
Overall Response Rate = 0.9								
0.900	0.900	0.01	0.02	0.03	0.01	0.02	0.05	0.00
0.890	0.910	0.02	0.03	0.04	0.03	0.04	0.07	0.03
0.875	0.925	0.03	0.04	0.06	0.05	0.06	0.10	0.08
0.865	0.935	0.04	0.05	0.07	0.06	0.08	0.12	0.12
0.850	0.950	0.05	0.06	0.08	0.08	0.10	0.15	0.17
Overall Response Rate = 0.8								
0.800	0.800	0.02	0.03	0.06	0.02	0.03	0.09	0.00
0.790	0.810	0.02	0.04	0.07	0.03	0.05	0.11	0.03
0.775	0.825	0.04	0.05	0.08	0.05	0.07	0.13	0.07
0.765	0.835	0.04	0.06	0.09	0.06	0.09	0.15	0.10
0.750	0.850	0.05	0.07	0.10	0.08	0.11	0.18	0.15
Overall Response Rate = 0.7								
0.700	0.700	0.02	0.05	0.08	0.03	0.05	0.13	0.00
0.690	0.710	0.03	0.05	0.09	0.04	0.06	0.15	0.03
0.675	0.725	0.04	0.07	0.10	0.06	0.09	0.17	0.07
0.665	0.735	0.05	0.07	0.11	0.07	0.10	0.19	0.10
0.650	0.750	0.06	0.09	0.13	0.09	0.12	0.21	0.15
Overall Response Rate = 0.6								
0.600	0.600	0.03	0.06	0.11	0.04	0.06	0.17	0.00
0.590	0.610	0.04	0.07	0.12	0.05	0.08	0.18	0.03
0.575	0.625	0.05	0.08	0.13	0.07	0.10	0.21	0.07
0.565	0.635	0.06	0.09	0.14	0.08	0.12	0.23	0.10
0.550	0.650	0.07	0.10	0.15	0.10	0.14	0.25	0.15

We explore various possible assumptions for α_t and α_c. To pick values of α_t and α_c for consideration, it is convenient to note that α^2 is equivalent to the *R*-squared from the regression shown in Equation (2)—that is, the proportion of the outcome variance that is explained by the propensity to respond. Therefore, we consider a range of values for α_t and α_c that yield a range of possible *R*-squared values in the treatment group (R_t^2) and control group (R_c^2):

- α_t = 0.27 and α_c = 0.22 (implying R_t^2 = 0.075 and R_c^2 = 0.05)
- α_t = 0.32 and α_c = 0.22 (implying R_t^2 = 0.10 and R_c^2 = 0.05)
- α_t = 0.39 and α_c = 0.22 (implying R_t^2 = 0.15 and R_c^2 = 0.05)
- α_t = 0.45 and α_c = 0.39 (implying R_t^2 = 0.20 and R_c^2 = 0.15)
- α_t = 0.55 and α_c = 0.45 (implying R_t^2 = 0.30 and R_c^2 = 0.20)
- α_t = 0.71 and α_c = 0.45 (implying R_t^2 = 0.50 and R_c^2 = 0.20)
- α_t = 1 and α_c = 1 (implying R_t^2 = 1 and R_c^2 = 1)

The key finding in Table 1 is that, given a set of assumptions regarding the correlation between outcomes and the propensity to respond, bias can be reduced by either increasing the overall response rate or reducing the differential response rate. For example, column 4 shows that an overall response rate of 60% yields a bias of no more than 0.05 only if the differential rate is 2 percentage points or less, but that if the overall rate is 90%, the differential rate can be as high as 5 percentage points.

3. Identifying Reasonable Values for Model Parameters

As shown in Table 1, the relationship between response rates and bias depends on the values of α_t and α_c. In order to determine which of these relationships should be used, we must identify which values for these parameters are reasonable to assume.

As the initial step toward identifying reasonable parameter assumptions, we first determine which of the assumptions in Table 1 are consistent with the previous WWC attrition standard.

Suppose that the previous standard had been developed to limit the possible bias to 0.05 standard deviations—the tolerance level for bias that we actually select for the new attrition standard, as discussed later. In this case, the values of α_t and α_c that are consistent with the previous standard are the ones for which an overall response rate of 80% and a differential response rate of 7 percentage points lead to a bias of 0.05. Using the row of Table 1 corresponding to these response rates (P_t = 0.765 and P_c = 0.835), we see that bias is approximately equal to 0.05 (that is, differs from 0.05 by no more than 0.01) in the first, second, and fourth columns, which correspond to α_t = 0.27 and α_c = 0.22 (column 1), α_t = 0.32 and α_c = 0.22 (column 2), and α_t = 0.45 and α_c = 0.39 (column 4). Therefore, those assumptions appear to be most consistent with the previous standard.

Our next objective is to determine whether the assumptions that are consistent with the previous standard (columns 1, 2, and 4 of Table 1) are also consistent with actual data from randomized trials in education. In fact, from existing studies, we could directly infer α in each group (treatment or control) if we could observe outcomes for both respondents and nonrespondents in that group. In each group, the attrition model implies a precise relationship between α and the relative outcomes of respondents and nonrespondents. Let Δ_g denote the difference in outcomes, in effect size units, between respondents and nonrespondents in group *g* (either the treatment [*t*] or control [*c*] group). It can be shown that

$$(5a)\ \ \Delta_g = E(y_g \mid z_g > z_g^*) - E(y_g \mid z_g \le z_g^*) = \frac{\alpha_g \times \phi(\Phi^{-1}(1-P_g))}{P_g(1-P_g)},$$

which implies that

$$(5b)\ \ \alpha_g = \frac{\Delta_g P_g(1-P_g)}{\phi(\Phi^{-1}(1-P_g))}.$$

Of course, we cannot observe outcomes for nonrespondents, so we cannot observe Δ_g directly. However, in studies that have both follow-up and baseline test scores, we can use the baseline test scores as proxies for the follow-up test scores because baseline scores are typically correlated with follow-up scores. Therefore, for each of several existing studies, we use the difference in *baseline* test scores between respondents and nonrespondents as the proxy for Δ_g, and we use Equation (5b) to calculate α_g for $g = t, c$.

Our data for these calculations come from three large-scale randomized trials conducted by Mathematica Policy Research for IES:

- Evaluation of the 21st Century Community Learning Centers
- Evaluation of Education Technologies in Reading and Mathematics
- Evaluation of Supplemental Reading Comprehension Interventions

One of these studies, the education technology study, had distinct interventions that were implemented in four different grade levels (first, fourth, sixth, and ninth), with random assignment occurring separately by grade level. Therefore, we calculated parameter values separately by grade level in this study.

For each study, Table 2 presents empirical values of key quantities (namely, the response rate [P] and the respondent–nonrespondent difference in baseline scores [Δ]) used to calculate α as well as the resulting value of α. Values are presented separately for the treatment and control groups. The studies generally had high response rates (of at least 80% in all but one case) in both the treatment and control groups. Effect size differences in baseline test scores between respondents and nonrespondents range widely across studies, from a low of 0.02 in the treatment group of the 21st Century evaluation to a high of 0.54 in the treatment group of the 6th grade education technology evaluation.

TABLE 2

RESPONSE RATES, RESPONDENT–NONRESPONDENT DIFFERENCES IN BASELINE TEST SCORES, AND IMPLIED CORRELATIONS BETWEEN TEST SCORES AND THE PROPENSITY TO RESPOND FROM THREE RANDOMIZED TRIALS

	Treatment Group			Control Group		
Evaluation	**P_t**	**Δ_t**	**Implied α_t**	**P_c**	**Δ_c**	**Implied α_c**
21st Century	0.81	0.02	0.01	0.83	0.10	0.06
Education Technology						
1st Grade	0.91	0.46	0.23	0.90	0.35	0.18
4th Grade	0.87	0.40	0.21	0.90	0.51	0.26
6th Grade	0.88	0.54	0.28	0.90	0.44	0.23
9th Grade	0.80	0.18	0.10	0.76	0.28	0.16
Reading Comprehension	0.89	0.31	0.16	0.88	0.32	0.17

For each study listed in Table 2, there are two empirical values of α —one for the treatment group and one for the control group. For the purposes of assessing whether specific assumptions in Table 1 are reasonable, it is sufficient to extract two characteristics of the values of α from each study: (1) the higher of the two values of α and (2) the difference between the higher and lower value of α. It is irrelevant whether the higher value of α in each study comes from the treatment or control group because, when generating the bias values in Table 1, we always assign the higher value of α to the group with the lower response rate (which, in Table 1, is arbitrarily chosen to be the treatment group). This approach is conservative because it assumes that treatment-control differences in α always reinforce (rather than mitigate) biases stemming from treatment-control differences in response rates.

Across these studies, the higher value in each pair of α's ranges from 0.06 to 0.28, with a mean of 0.19. In addition, the difference between the higher and lower value of α ranges from 0.01 to 0.06, with a mean of 0.05. Of the various sets of assumptions for α_t and α_c represented in the different columns of Table 1, the assumptions in the first column of Table 1 appear to approximate the empirical values of α most closely.

In summary, data from the three studies are consistent with low values for the correlation between test scores and the propensity to respond, as well as with small differences in that correlation between the treatment and control groups. As discussed earlier, correlations in the lower part of the range explored in Table 1—namely, the correlations represented by the first, second, and fourth columns of the table—are also consistent with the previous attrition standard. Taken together, these findings suggest that values of α_t and α_c in the range of those shown in the first four columns of Table 1 are reasonable assumptions. In fact, the data used in our analysis lean toward the first column. However, for certain populations of students not included in those studies, such as older students who volunteer to participate in a dropout prevention program, attrition may be more correlated with the outcome, in which case more conservative assumptions (such as those in the fourth column of Table 1) would be justified.

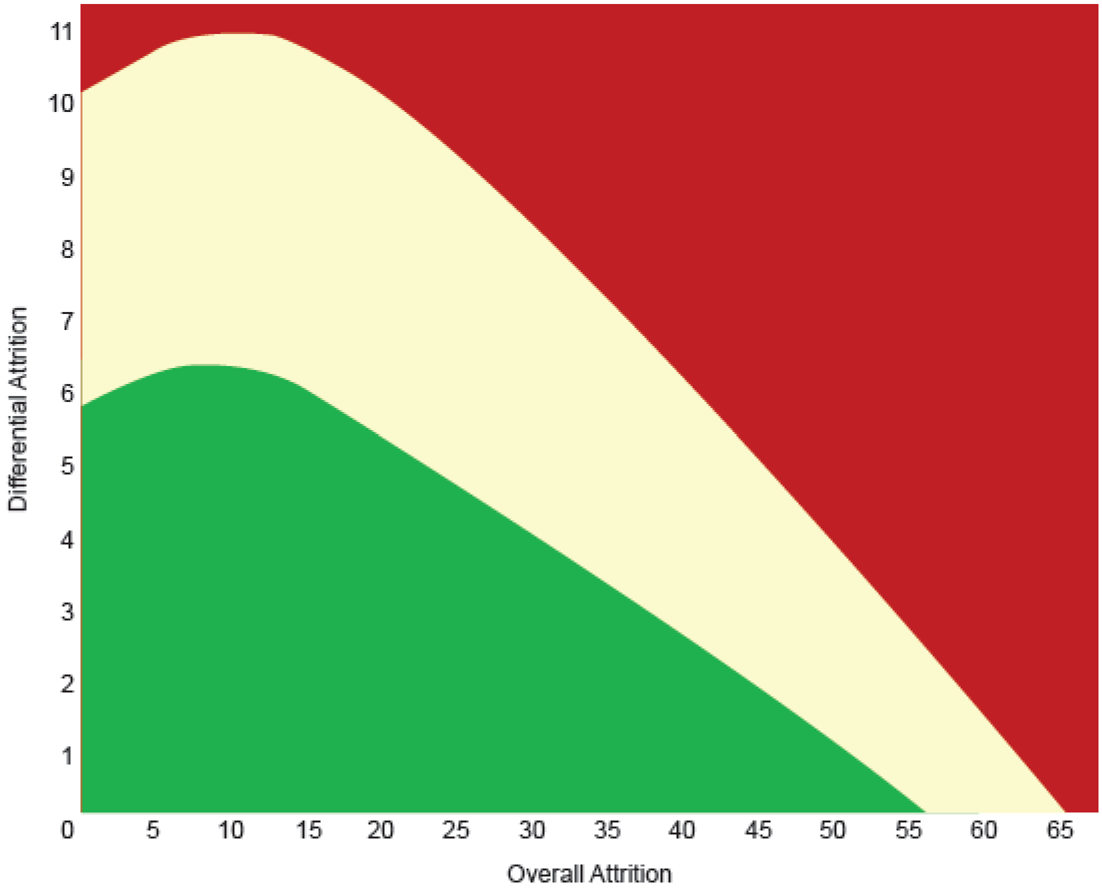
4. Attrition Tradeoffs Assuming a Constant Bias

The preceding findings enable us to map out a specific relationship between bias and response rates by setting α_t and α_c equal to reasonable values. That is, rather than having to consider the wide range of possible relationships between bias and response rates shown in Table 1, we can now select a particular relationship that corresponds to our chosen values of α_t and α_c.

For the purposes of developing an attrition standard, we choose a threshold degree of tolerable bias. Using a selected relationship between bias and response rates from Table 1, we then calculate which combinations of overall and differential response imply biases that exceed or fall below the threshold. This approach highlights the tradeoff between overall and differential attrition and can be illustrated graphically. Figure 1 uses a bias threshold of 0.05 standard deviations of the outcome measure. The green/bottom-left region shows combinations of overall and differential attrition that yield attrition bias less than 0.05 under pessimistic (but still reasonable) assumptions ($\alpha_t = 0.45$ and $\alpha_c = 0.39$), the yellow/middle region shows additional combinations that yield attrition bias less than 0.05 under optimistic assumptions ($\alpha_t = 0.27$ and $\alpha_c = 0.22$), and the red/top-right region shows combinations that yield bias greater than 0.05 even under optimistic assumptions.

FIGURE 1

TRADEOFFS BETWEEN OVERALL AND DIFFERENTIAL ATTRITION



To get some indication of how large the relative bias is, note that for a nationally normed test, a difference of 0.05 represents about 2 percentile points for a student at the 50th percentile. For example, if the reported effect suggests the intervention will move the student from the 50th percentile to the 60th percentile (a 0.25 effect size), the true effect may be to move the student from the 50th percentile to the 58th percentile (a 0.20 effect size). Doubling the tolerable bias to 0.10 means that an intervention that reportedly moves a student from the 50th percentile to the 60th percentile may move the student only to the 56th percentile—a scenario that seems to imply a fairly large bias. With these considerations, we set the threshold degree of tolerable bias to be 0.05.

5. Using the Attrition Bias Model to Create a Standard

In developing the topic area review protocol, the principal investigator (PI) considers the types of samples and likely relationship between attrition and student outcomes for studies in the topic area. When a PI has reason to believe that much of the attrition is exogenous—for example, parent mobility with young children—more optimistic assumptions regarding the relationship between attrition and outcomes might be appropriate. On the other hand, when a PI has reason to believe that more of the attrition could be endogenous—for example, high school students choosing whether to participate in an intervention—more conservative assumptions may be appropriate. The combinations of overall and differential attrition that are acceptable given either optimistic or conservative assumptions are illustrated in Figure 1, and translate into evidence standards ratings:

- For a study in the green/bottom-left region, attrition is expected to result in an acceptable level of bias even under conservative assumptions, which yields a rating of *Meets Evidence Standards*.

- For a study in the red/top-right region, attrition is expected to result in an unacceptable level of bias even under optimistic assumptions, and the study can receive a rating no higher than *Meets Evidence Standards with Reservations*, provided it establishes baseline equivalence of the analysis sample.
- For a study in the yellow/middle region, the PI's judgment about the sources of attrition for the topic area determines whether a study *Meets Evidence Standards*. If a PI believes that optimistic assumptions are appropriate for the topic area, then a study that falls in this range is treated as if it were in the green/bottom-left region. If a PI believes that conservative assumptions are appropriate, then a study that falls in this range is treated as if it were in the red area. A PI chooses whether the optimistic or conservative assumption applies for the review. However, once the assumption is chosen, it will be applied to all studies reviewed in that area and not vary across studies.

When the unit of assignment differs from the unit of analysis and both types of units have the potential to leave the study, the attrition standard will be applied separately to (1) all units of assignment and (2) the units of analysis contained within any units of assignment that did not leave the study before follow-up data collection. The study must meet the attrition standard for both the units of assignment and the units of analysis in order to *Meet Evidence Standards*.