**Supplemental Digital Content 3**A System for Rapidly yet Rigorously Evaluating the Quality of Randomized Controlled Trials
*Journal of Clinical Psychopharmacology*
Eric G. Smith and Hannah L. Grigorian

**Additional Information about the BIS FOES system Intended to Aid the Use or Teaching of the System**

Below we include additional information about the BIS FOES system. This information includes rationales for choices made in designing the system (e.g., specific benchmarks chosen), potential for refinement to the BIS FOES system, approaches to help remember the BIS FOES acronym and criteria, thoughts about how the BIS FOES system might contribute to efforts to improve RCT Abstracts or Limitations subsections, and further background information about some of the BIS FOES criteria. Readers should feel free to skip this Supplement if they feel the main article has provided all the information they need, or to pick and choose among the topics in this Supplement that might interest them.

**Rationale for Benchmarks for Baseline Characteristic Balance, Loss During Follow-up, and Study Size**

Our Tutorial was intended to give a relatively concise overview of the entire BIS FOES system. As such, we provided the benchmarks that we use to evaluate certain criteria, but did not explain the basis for those benchmarks. Those rationales may be useful for some users of the system, so we explain them below:

***Size/Sufficiency of Randomization Benchmark of 10+% Difference in Baseline Characteristics as indicating a “Substantial Imbalance”***

We use the benchmark of a 10% difference as indicating a “substantial” difference between baseline characteristics. This benchmark is based on an accepted, commonly-used metric for randomized and non-randomized studies (e.g., studies using propensity score matching). This traditional metric evaluates whether all baseline characteristics are balanced between study arms within a standardized difference of <0.1.1, 2 However, in our experience, standardized differences (a measure of differences in prevalence which includes the standard deviation) are almost never provided in Baseline Characteristics Tables. Instead, overall prevalences (in the form of the percentage of participants in each study arm sample who possess that characteristic) are almost always provided. Thus, for simplicity, and because a case can be made that differences of <10% between study arms in the prevalence of any given baseline characteristics are unlikely to substantially bias a RCT’s findings unless that characteristic is a particularly strong risk factor, we use the benchmark of a 10% difference in overall prevalence between study arms to assess the sufficiency of randomization.

As an illustration, simple subtraction of the reported percent prevalences in each study arm can be done to ascertain if the differences in prevalence of that characteristic between the study arms are less than 10% or not. For instance, if 50% of the participants are female in one study arm and 42% are female in the other, 50-42% = 8%, which would be a <10% difference).

***Loss During Follow-up Benchmark of 20+% Loss During Follow-up as Particularly Concerning***

Our benchmark of 20% is based on a simplification of the benchmark discussed by Schulz and Grimes in 2002,3 citing earlier work by Sackett and colleagues. Schulz and Grimes discuss a three-part criteria, with less than 5% attrition representing a fairly minimal concern, five percent to less than 20% attrition representing significant attrition, and 20% attrition representing attrition that would constitute a serious concern.3 As we point out in the Tutorial, losses during follow-up of ≥20% are common in mental health RCTs. Therefore, since one benchmark is easier to remember than two, we typically teach just the 20% benchmark to trainees. However, when teaching trainees who are reviewing a manuscript with particularly low attrition, we will usually point out at the time if the RCT features an attrition rate that is notably low (e.g., <5%) for a mental health RCT.

***Size/Sufficiency of Randomization Benchmark for Small, Medium, and Large Study Sizes***

Benchmarks for judging “small”, “medium”, and “large” RCTs are less universal. We chose a benchmark of ≤30 patients per arm to define a “small” RCT based on the statistical consideration that the number 30 is the approximate point when a t distribution begins to better approximate a standard normal, or Z, distribution.4 The benchmark for “large” RCTs was initially based on anecdotal statistical guidance that studies with 100+ patients per arm were generally of the most value. This precise benchmark, however, has been used to distinguish “small” from “large” RCTs in at least one study evaluating the potential bias in RCT findings that is associated with study size.5 These benchmarks left the middle range of 31-99 patients per arm for what we consider “medium” RCTs.

 Any benchmarks used to judge study size (and the accompanying general likelihood that randomization worked sufficiently to balance the study arms in baseline characteristics) are intrinsically somewhat artificial. In general, a continuum exists in which the larger the RCT, the more likely randomization will have sufficiently balanced the baseline characteristics. However, one of the most detailed studies of the association between RCT study size and variations in effect size provides some support for our size categories.6 This study found that, compared with trials of 1000 patients or more, treatment effects were, on average, 48% larger in trials with < 50 patients (presumable < 25 participants/arm)), 30-34% larger in trials with 50-199 participants (presumably 25-<100 participants per arm, and 10-19% larger in trials with 200-999 participants (presumably 100-<500 participants/arm).6 (Thus, their smallest size classification is close to 30 participants/arm, and the next two categories, the analogues to our “medium” and “large” studies, used benchmarks similar to ours and showed an increasing amount of agreement between their treatment effect estimates and those from the largest category). This study’s categorization scheme does suggest that there may be value in considering a 4th category denoting “very large” studies with 500+ participants per arm.

Also worth considering is the fact that our size benchmarks have demonstrated their value repeatedly to us. As mentioned in the Tutorial, we have consistently observed in our journal club that these categories are fairly predictive of the number of substantial imbalances that we are likely to observe in the RCT’s Baseline Characteristics Table. Nevertheless, alternative benchmarks certainly could be considered.

We chose these benchmarks initially to help delineate the expected impact of study size on the sufficiency of randomization obtained in the RCT. However, it is important to note that it is not clear if the impact of study size on the sufficiency of randomization is in fact the most important factor underlying the associations that have been repeatedly observed between study size and treatment effect. Multiple mechanisms for this “small study effect”7 have been hypothesized, ranging from publication bias (the tendency for smaller studies to be published only if they have significant results and/or larger effect sizes), increased likelihood that outcomes will be selectively reported, and possibly more methodological flaws.6, 8 However, it is also possible that the larger effects at lower sample size reflect bias do not always reflect bias. Larger RCTs may have greater heterogeneity in participants or more variation in, or less oversight over, implementation of the intervention,6, 8 factors which might be expected to lower the effect size.

At a minimum, study size is an important criteria to assess for RCTs because it affects the sufficiency of randomization and potentially other factors that may bias a RCTs findings. In addition, our size benchmarks have some support in the literature, but we also consider them open to refinement.

In summary, all 3 of these benchmarks discussed above seek to impose a structure on phenomenon whose impact on RCT quality is likely on a continuum. This is done to alert users of the BIS FOES system to the circumstances in which they should pay particular attention to, and potentially have heightened concern about, these criteria. As such, these benchmarks are intrinsically somewhat arbitrary, but we have found them useful in our journal club. Users can modify these benchmarks if they feel strongly that this change would make the BIS FOES system better suit their purposes.

**Potential for Refinements to the BIS FOES system**

 We feel that the BIS FOES system as outlined in the Tutorial is extremely useful in its present form, and strikes an important and “user-friendly” balance between convenience, memorability, and rigor. Nevertheless, as hopefully our discussion of the benchmarks immediately above makes clear, we hope that the BIS FOES system can be seen as something of an “open source” system. As such, we see it as potentially open to modification by others for their purposes, and we welcome suggestions for refinements to the system.

 In our view, however, the most important aspect of the BIS FOES system to preserve through any modifications is, if possible, its relatively simple 7 criteria framework. Such a framework raises the possibility that a number of users can learn to implement the system *without needing to necessarily resort to a written checklist*. The benefits to this are obvious, in that it would allow for relatively rigorous reviews of RCTs to occur in almost any setting. So we ask users seeking to modify the system to consider looking first to modifying the structures within the criteria, such as benchmarks, subcriteria, and/or criteria content (e.g., including more about allocation concealment within “Blinding”), rather than the criteria themselves, unless they feel they have very strong reasons to do so.

In preparing this Tutorial and Supplements, we ourselves have noted areas for possible improvement or refinement. One example would be including within Special Considerations a specified subcriteria to note any deviations from the original study protocol. (If desired, if the practice of naming these items starting with “S” is considered desirable, a subcriteria noting the “Similarity [of the RCT] to the Original Protocol” could be added). A second example that would entail a more extensive change to the BIS FOES system would involve having users assess whether the outcome of greatest interest (usually the primary outcome) is subjective or objective as the very first RCT element evaluated in the review. Then, if the outcome is objective, users could flip the order of the first 3 BIS FOES criteria to be “SIB FOES” rather than “BIS FOES.” The idea behind this is to serve as a reminder that Blinding may be considerably less important when outcomes are objective, and thus when objective outcomes are used, the acronym might benefit from listing Size first since it may have such a major impact on the RCT’s quality. The acronym could even be changed to “SO (BIS or SIB) FOES?” (where “SO represents “Subjective/Objective” and the question mark indicates one is deciding whether to conduct a BIS FOES or SIB FOES review). Adding flexibility to a system can add complexity, however, and we suspect this approach is too complicated for most trainees. However, certain users who review a lot of RCTs with subjective outcomes and a lot of RCTs with objective outcomes may want to experiment with this approach. It is possible that it could help them remember the significance of the type of outcome measure on the importance of blinding.

**Different Memory Aids for the BIS FOES system**

 A key element in the convenience of the BIS FOES criteria in our mind is whether users can ultimately commit at least the basic 7 criteria of the system to memory (and ideally some or all of the subcriteria as well, with the possible exception of the detailed list of specific “Special Considerations”). Below we suggest several possible memory aids for users to help them remember what criteria the BIS FOES acronym represents, and potentially which criteria have multiple subcriteria.

***One Suggestion for Visualizing the BIS FOES acronym***

Different individuals learn in different ways. Some individuals may find it helpful to think of the “BIS FOES” acronym as representing a description of business rivals competing for customers. (The term “BIS” [also spelled as “BIZ”] is sometimes use as a slang term for business, and “rivals” can be seen as another term for “FOES”). This imagery, although not perfect (the imagery does not relate directly to the evaluation of randomized trials, nor is it as dramatic or humorous as some other mnemonics), may help some users remember the overall acronym. Remembering the overall acronym may then help at least some of these readers better remember the seven criteria of the system itself.

***Three Potentially Helpful Aspects to the Ordering of the BIS FOES criteria***

There are other strategies besides just the acronym and the imagery mentioned above that might help some users to remember the BIS FOES criteria. We recognize fully remembering a system with 7 criteria may be challenging, especially when first learning the system. Therefore we point out below that the ordering of the BIS FOES criteria has qualities that might be helpful in remembering the acronym letters and what these letters represent:

**1) One can conceptualize the first 3 criteria, Blinding, Intent-to-Treat, and Size (Sufficiency of Randomization) as relating to initial (or “first”) study activities, and the second three criteria (Follow-up [Loss During], Outcomes, and Effects), which come “later” in the acronym, as relating to later study activities (at least with respect to participants progressing through the trial).**

This is obviously an approximation of sorts. However, for instance, when blinding occurs, the implementation of Blinding of at least participants and/or clinicians typically occurs either immediately after randomization or shortly thereafter. Intent-to-Treat analysis, the second BIS FOES criteria, relates to the question of whether an intent-to-treat sample is analyzed, and by definition a full intent to treat study analyzes everyone initially randomized. Finally, the Size of the study relates to the number of participants randomized to each study arm, and randomization essentially is what starts each participant into the randomized portion of the RCT. Thus, each of these first 3 criteria can be thought of, to some extent, relating to activities occurring at or before the start of the follow-up period for an individual participant. In contrast, the next 3 BIS FOES criteria ([Loss During] Follow-up, Outcomes, and Effects) can be thought of, to some degree, as pertaining to events occurring *after* a participants start of the study. Participants need to be participating in the follow-up period to subsequently be “lost” during that period, and while the outcomes of the study are almost always pre-specified, the outcomes that are counted in a RCT are those outcomes that occur after randomization. Finally, the effects observed in a study refer to the aggregation of all the outcomes observed during follow-up.

We present this conceptualization of the BIS FOES criteria as relating to initial and subsequent aspects of the RCT because we suspect this distinction, approximate though this division among criteria may be, may be useful in helping some users to remember what the BIS FOES acronym’s letters represent.

**2) The criteria are listed in approximate order of ease of assessment**

The first 3 criteria typically have no formal subcriteria (Intent-to-Treat) or only one subcriteria (determining if an assessment of blinding was done for Blinding, and assessing whether substantial imbalances exist in the randomized sample for Size/Sufficiency of Blinding). In contrast the next 3 criteria have multiple subcriteria to assess, and the 6th criteria, Effects, may have some calculations to actually perform, which adds complexity even though these calculations are usually relatively quick to perform. The 7th criteria, Special Considerations, can be used as a general “catch all” to comment upon other aspects of the study, but if one wants to be more systematic has nine subcriteria that could be examined (SDC-2). Thus, an approximate ordering to the BIS FOES criteria exists in which one proceeds from more easily assessed criteria to criteria with more subcriteria as one progresses through the acronym.

Remembering this pattern may help some users remember what the letter represent and help jog their memory that there are multiple subcriteria to consider for the later criteria.

**3) If it helps some users, the BIS FOES system can even be conceptualized as being a 5- or 6-criteria system rather than a 7-criteria system**

The more elements that one has to remember, generally the more likelihood that one or more of them will be forgotten, and it can be argued that 7 criteria are a lot to remember. We have tried to adapt to the reality of having a 7 criteria system by creating the “BIS FOES” acronym to help users remember the key criteria. Still, even with the acronym, some users, or some trainees , may find it hard to remember all 7 elements. Some users may find it more helpful to think of the system as a 6-criteria system, with the 7th criteria simply serving as a placeholder for other impressions about the RCT’s quality. Other users may even prefer to think of it as a 5-criteria system, with the first 5 criteria representing “quality” indicators for RCTs, while the sixth criteria, Effects, is seen as more of a guide to the actual results of the trial and their importance, rather than a quality indicator *per se.* The 7th criteria, as described above, is unusual in that it is a “placeholder” or “catch-all” criteria, and therefore could be considered as separate from the first 5 criteria as well. Thus, a nice feature of the BIS FOES criteria is that the two atypical criteria appear at the end of the acronym, allowing the possibility of thinking of it as a simpler 5-criteria system with 2 additional, slightly different elements at the end.

In summary, we point out these potential learning aids for the BIS FOES system because we recognize that learners differ greatly in how they best learn material. Users should feel free to use any or all of the approaches above if helpful.

**Could BIS FOES (and other RCT evaluation systems) help improve RCT Abstracts and Limitations Subsections?**

We hope the BIS FOES system, which we have found definitely helpful in reviewing RCTs, might potentially contribute (along with other RCT evaluation systems), to ongoing efforts to enhance the content of RCT Abstracts and the Limitations subsections in RCT manuscripts.

Regarding Abstracts, as pointed out in the Tutorial, we find that RCT Abstracts often are lacking information about several key BIS FOES criteria. In particular, we find that information is frequently lacking concerning the BIS FOES criteria of (Loss During) Follow-up, whether Intent-to-Treat analyses were conducted, and some study Effects (specifically, safety findings and sometimes effect size measures). Potentially, journal editors or reviewers might even consider asking RCT authors to also report the number of characteristics that are substantially imbalanced between the study arms (perhaps as a fraction of the total number of baseline characteristics evaluated, to not “penalize” those studies listing many baseline characteristics in their Baseline Characteristics Table). Alternatively, but possibly less informatively, authors could be asked to report the average difference observed between all the baseline characteristics. We can only recall one Abstract we have read that provided information of this type, and that was for a single characteristic (the baseline score of the primary outcome).9

We even think Journal editors should consider expanding the length of Abstracts, if needed and if consistent with their format, to allow the inclusion of more information concerning indicators of potential trial quality. We also recognize that other evaluation systems may have evidence to support another RCT element as being a higher priority to include in Abstracts. Regardless of whether any added RCT elements come from BIS FOES or another system, we would applaud efforts to make RCT Abstracts more informative. The basic guiding principle, in our view, should be to make the RCT Abstract as informative as possible. This would allow busy readers to be able to quickly yet thoroughly assess how much additional time they should devote to the manuscript.

Regarding Limitations subsections, we find that important limitations to a RCT often go uncommented upon in this subsection, while other seemingly less important limitations are noted. These important limitations include such considerations as a lack of blinding or the possibility that blinding was not completely effective, the lack of a full intent-to-treat analysis, small size, substantial imbalances in baseline characteristics, and the presence of multiple comparisons. While we agree that Limitations subsection of the Discussion is the ideal location for authors to comment on limitations that only they may be aware of, we also feel strongly that it is the best place in the manuscript to offer an overall summary or recap of all of an RCT’s important limitations. In our view, many Limitations section would be improved if authors were to address more of the issues that are highlighted by BIS FOES (or other systems for evaluating randomized trials) that are particularly pertinent their trial. Hopefully, study authors who are aware of the BIS FOES system will use it and/or other RCT evaluation systems to help guide their writing of their Limitations subsections. When this does not occur, we hope journal editors and reviewers encourage authors include more of the information highlighted by BIS FOES or other systems for RCT evaluation in their Limitations subsections.

**Additional Considerations / Teaching Notes Concerning Specific BIS FOES Criteria**

This section presents some additional background that may be useful for some educators to help them in teaching the system, or that may be of interest to some users of the system. Some of the material is quite detailed, but in these cases we strive to present the simplest information first in the section. This material is not essential for understanding the BIS FOES system, and readers should feel free to skip this section, scan this section, or read it in depth as they prefer.

**Blinding**

Of interest, the overall importance of blinding in limiting bias in RCTs has been questioned recently. One prominent study from 2020 has called into question the importance of blinding10 and cites other similar studies as supporting this conclusion.11, 12 One of these “supporting” studies did show that a lack of or unclear blinding of outcome assessors was associated with a negligible effect on treatment effect estimates when results for studies with subjective and objective outcomes were pooled. However, the same study also showed that in studies with subjective outcomes, lack of or unclear double blinding, or lack of/unclear blinding of outcome assessors, was associated with a 23% or greater exaggeration of the treatment effect.11 In addition, even the 2020 study acknowledged that the most rigorous designs for answering this question showed a lack of blinding produced large biases.13, 14 These particularly rigorous designs feature either within-trial comparisons of blinded and nonblinded outcome assessment, or trials which randomized patients to blinded or unblinded substudies. The 2020 paper notes these more rigorous designs especially add to the evidence that blinding is particularly important for trials with subjective outcomes. These more rigorous studies find that nonblinded RCTs with subjective outcomes report 36% larger treatment effects on average than blinded RCTs with subjective outcomes.10 The authors even state in their concluding sentence that blinding should remain an important “methodological safeguard” for RCTs when it is feasible. For these reasons we continue to put great value on whether a study was blinded in BIS FOES and in our RCT assessments (especially for mental health RCTs, which generally use subjective outcomes).

Also of interest, the importance of the “guess test” has also been disputed.15, 16 The 2001 CONSORT guidelines state that the guess test assessment should work in principle, but then simply states that, in practice, participants successfully guessing their assignment does not mean that blinding was unsuccessful.15 No further detail is added. However, the CONSORT statement (citing a prior commentary by mental health researchers16) does provide an argument for why clinicians specifically might be able to accurately guess a participant’s assignment even if no obvious problems with the blinding exist. This can occur because clinicians may have expectations that in a placebo-controlled trial, participants getting active treatment are more likely to experience a response. If more participants did in fact experience a response who were receiving active treatment, and the clinicians simply guessed that all most or all participants experiencing a response were on active treatment, this could bias their responses on the guess test away from chance.15, 16 To us, it seems that the same argument could be made for participant responses to the guess test as well, although that point is not made in either reference. In essence, the argument is that the guess test can indicate problems with blinding, but it can also indicate after-the-fact expectations related to assumptions that those participants doing better are on active treatment, rather than anything that interfered with the study (such as problems with blinding) while it was underway

We note that examples do exist of RCTs in which significant benefits for active treatment were reported compared to the control (at least for one of the two main outcomes), but participants were unable to accurately guess their treatment assignment.17 In this study, it seems to us that the guess test result has value, since it might be seen as strengthening the evidence for the observed effect in this specific evidence by suggesting the observed response is attributable neither to problems with blinding nor expectations that the patients who did better were on active treatment. (In this specific study, the participants were diagnosed with a condition that might have interfered with their ability to recognize that they were doing better in specific circumstances). Thus, even if one accepts the concerns discussed above, the guess test might still have value when it is “negative” (that is, suggesting that study agents could not infer successfully what treatment participants are receiving). However, how often this will occur if an intervention produces notable benefits for participants is unclear. Obviously, more research would be helpful to better delineate under what circumstances the guess test helps inform inferences about the integrity of blinding.

In summary, it would appear that the guess test can suggest that problems with blinding might have affected the results of an RCT, but not establish that this is the case. Because of the potential for some mental health trials to show a very high rate of participants guessing their assignment (up to 94% of participants, as we note in the Tutorial), we still feel that the guess test is a useful assessment to consider when evaluating blinding as a “bounds-setting exercise.” In our view, the guess test can provide an approximate measure of the potential magnitude of problems with blinding, even though deviation from chance in guesses of treatment assignment could be from other reasons as well. If, however, an investigator chooses not to administer a guess test because of the complexities of its interpretation or for other reasons, this should not be seen as a potential quality issue for the RCT.

Finally, of note, some authors have pointed out that even seemingly objective outcomes such as mortality can be influenced, at least to a slight extent, by a lack of blinding.18 These authors recommend blinding of participants and study personnel when possible even for mortality outcomes (and presumably for other objective outcomes as well).

**Intent-to-Treat**

As indicated in the Tutorial, the authors of a 2015 metanalysis indicated that deviations from a full intent-to-treat analysis could lead to substantial bias.19 These authors also decried the inconsistent terminology used for modified intent-to-treat analyses, which ranged from terms as confusing as the “intent-to-treat subset,” or “patients evaluable for efficacy.”19

We would like to concur with these concerns, and point out that there is inconsistency even when the term “modified intent-to-treat” is used. Some writers view modified intent-to-treat analyses as excluding participants who are randomized but fail to receive even one dose of the study medication.20 However, in our reading of mental health RCTs, we find the more common use of the term to refer to the practice of excluding individuals who lack any post-baseline assessments of the outcome of interest. In fact, the 2015 meta-analysis itself found that this approach accounted for 47% of all “modified” intent-to-treat analyses.19 Occasionally, there are other examples in which the definition of “modified” intent-to-treat analysis is stretched even further.

 We suggest that some consideration be given to adopting more descriptive terminology, such as using terms like “recipient” ITT (possibly abbreviated “rITT”) for analyses of all individuals to receiving 1 or more doses of the study intervention, and “post-baseline-assessed ITT” (possibly abbreviated “pbaITT”) for analyses evaluating all participants receiving at least one post-baseline assessments. (In general, however, we favor efforts to retain the entire intent-to-treat sample by imputing or otherwise estimating results for participants who only received baseline assessments).

 For analyses of any other restrictions to the sample besides those above, we suggest perhaps that the term intent-to-treat in any form should be avoided entirely. Instead, a term like “restricted sample analysis” or something similar could be used instead to refer to those circumstances in which the sample was restricted beyond simply excluding those who did not receive any doses of the intervention, or any post-baseline assessment, but not so restricted that only “completers” were evaluated. In our opinion, this more diverse yet systematized terminology would better alert the reader (potentially reminding them each time they read the term) to how a particular RCT’s analyses may have diverged from a full ITT analysis.

**Size** (Sufficiency of Randomization)

First, it should be noted that Baseline Characteristics Tables typically do not provide direct information allowing one to easily assess whether 10+% differences exist in ordinal/continuous variables (such as rating scale scores, vital signs, or lab values). Fortunately, is it fairly easy to determine if a 10+% difference exists through use of a simple mathematical “trick”. To determine what value is 10% higher or lower than the mean value for a score or measure provided in the Table, simply “move” the decimal point one place to the left and add or subtract this value to the mean value. For example, a 10% difference from a mean depression rating scale of 22.0 would require at least a difference of either + or – 2.20 (2.20 is the value obtained from the value 22.0 after the decimal point is moved one place to the left. Thus, if the rating scale score in the other study arm(s) was 24.2 or greater, or 19.8 or lower (22.0 + 2.2 and 22.0 – 2.2, respectively), the imbalance in that rating scale score would be 10% or greater.

 When reviewing which patient characteristics show substantial imbalances in the Baseline Characteristics Table with trainees, we often will note to trainees that it is likely that “not all baseline characteristics are created equal.” When certain characteristics such as the baseline score on a primary or secondary outcome measure, or characteristics such as illness history or degree of current treatment for the illness, are imbalanced, these factors may be of particular concern for having the potential to bias the RCT’s findings . This concern arises from the fact these characteristics are often directly related to the condition and the outcome being studied.

 There can even be baseline characteristics for which imbalances < 10% can add substantial bias as well, if they are a particularly strong risk factor. However, as pointed out in our section on benchmarks above, the 10+% benchmark is based on a simplification of a generally-accepted research practice, and we find it to be a good approach for trainees to use to assess the sufficiency of randomization. .

We also often do ask trainees to make an initial assessment, based on the the total set of substantial imbalances seen, to make an assessment whether the imbalances remaining after randomization appear to have the potential to bias RCT findings in *favor* of the intervention of interest or *against* the intervention of interest. If imbalances are present, the case can be made that it is less damaging to the study if they appear to bias *against* finding a significant result for the intervention of interest being studied. (In this circumstance, any significant finding favoring the intervention would be occurring *in spite of* the suspected bias, rather than the bias operating to increase the chance of a false association being observed between the intervention and the outcome). Nevertheless, it should be remembered that imbalances in measured characteristics suggest that at least some imbalances in unmeasured characteristics are also possible, and it is impossible to know how strong or weak a risk factor these potentially imbalanced unmeasured characteristic might be. For this reason, the most preferred circumstance is when the randomization can be performed that is expected to work very well based on theoretical grounds (e.g., the randomization is performed on a large sample) and that randomization results in few or no observed substantial imbalances being observed between the study arms).

 We also note that some investigators will choose to include in regression analyses those baseline characteristics that are still judged to be substantially imbalanced after randomization. This is done by some investigators as a means to attempt to reduce the impact of those imbalances on the RCT’s findings. Although not as desirable an approach as not having the imbalances in the randomized sample at all (since any regression approach requires modelling assumptions), usually this regression approach is seen as preferable to not addressing these imbalances in any manner. However, this approach still does not allow for inclusion of any unmeasured baseline characteristics, so it is still expected to be inferior to a randomization that would result in fewer or no imbalances.

**(Loss During)** **Follow-up**

To clear up one potential source of confusion, we use the term “loss *during* follow-up”, rather than “loss *to* follow-up”, because we have found this terminology less confusing to trainees. Some Study Flowcharts use the term “loss to follow-up” to indicate only a particular subset of the total participants who attrition from the study. Presumably this narrowly-defined “loss to follow-up” term is given to participants who lose contact with the study team as opposed to those who are removed by the clinical team or who have other reasons for attrition. Because we are interested in the potential biasing effect of total attrition in each of the study arms, we use the term “loss during follow-up” instead to avoid confusion.

Of interest, Schulz and Grimes actually mention in a 2002 paper that some journals will refuse to publish trials with ≥20% attrition.3 However, a strict, rigid rule in this regard could be seen as undesirable since it would seemingly disincentivize investigators from conducting studies with long follow-up periods or in patient populations in which high loss during follow-up might be expected. As we note in the Tutorial, and Schulz and Grimes also note, concerns about loss during follow-up are increased when outcomes are infrequent.3 One way to visualize this concern is that if there are relatively few outcomes observed in the sample overall, then it takes only a few additional outcomes to have occurred within the participants lost during follow-up, but not observed (and thus not counted) because the participants were lost during follow-up, to potentially substantially alter the RCT’s results.

Guyatt and colleagues21 have proposed a slightly more involved approach to assessing the potential impact of loss to follow-up that captures the importance of whether outcomes are infrequent or frequent. This approach involves deliberately constructing a “worst case” scenario in which every person lost during follow-up in one study arm is assumed to have experienced an outcome, and every person lost during follow-up in the other study arms(s) is assumed not to have experienced the outcome In the example they provide, even a 3% attrition rate could completely eliminate an observed large treatment effect of a 0.5 relative risk reduction *if the rate of observed outcomes is quite low* (3% in the treatment group and 6% in the control group). This elimination of the treatment effect occurs because in their example of a 1000 patient sample, a 3% outcome rate and 3% loss during follow-up rate both total 30 individuals. If one assumes that all 30 of the individuals lost during follow-up from the treatment group experienced the outcome (in this example, the outcome was mortality), and none of those lost during follow-up in the control group did, the result is that the treatment group and control group would have been observed as having experienced 60 outcomes (deaths) if loss during follow-up *had not occurred*. (That is, 30 observed deaths + the 30 deaths lost during follow-up in the treatment arm [that we are now assuming were observed, since we are assessing what would be observed if no loss during follow-up occurred], and 60 observed deaths + 0 deaths lost during follow-up in the control arm, since we are assuming no deaths occurred in those individuals lost during follow-up). To put it another way, in this “worst case” scenario, it is possible that the entire treatment effect that was observed (the 0.5 relative risk reduction) was actually due to bias arising from loss during follow-up. In this case, the treatment in actuality would be associated with absolutely no benefit.

Guyatt and colleagues provide a second example in which the mortality outcomes occurred at a much higher frequency (20% in the treatment group [200 deaths] and 40% in the control group [400 deaths]). If 3% attrition is observed, then even assuming the “worst case” scenario has relatively little impact on the treatment effect estimated by the study. The treatment group would now have 200+30 deaths and the control group 400+0 deaths, a much lesser change in the findings (the relative risk reduction would decrease, but only modestly, to 0.43, rather than completely disappearing as in the infrequent outcome scenario). 21

There appear to be several implications of this exercise: First, it strongly reinforces the basic point that the more infrequent an RCT’s outcomes are, the greater the potential for any given amount of loss during follow-up to create substantial bias in the RCT’s findings. Second, there are actually some circumstances in which study attrition can be noted, but be judged as being unable to substantially bias a study’s findings (even in the “worst case”). However, we suspect these instances to be pretty infrequent, or even rare. In the Guyatt and colleagues approach, attrition had to be very low (just 3%) and the outcome had to be rather common (20 to 40% [on average, 30%]) or *10X* the rate of loss during follow-up [3%] that was observed), for attrition to have only modest impact even in the “worst case” scenario. Third, the infrequent outcome example that they provide should teach us to have great concern when loss during follow-up rates are close to the outcome rates in the study that during follow-up could substantially bias the RCT’s results.

Although Guyatt and colleagues indicate that they favor this approach to evaluating loss during follow-up, we do not teach the routine use of this approach. Of note, this approach has limitations in that it can only be applied to studies with dichotomous (present versus absent) outcomes, not outcomes such as changes in rating scales. Second, it is unclear how likely the worst case scenario is to ever occur. This approach can make a genuine contribution in identifying those few cases in which loss during follow-up seems unable to substantially bias the study results, even in the “worst case scenarios.” However, these circumstances seem particularly unlikely to occur in mental health RCTs, which typically have higher attrition rates. We find that, for teaching purposes, the more widely applicable 20+% attrition benchmark, which others have also suggested,3 to be more useful. (We consider this benchmark to be more widely applicable because it can be applied in studies of either continuous or dichotomous outcomes, unlike the Guyatt and colleagues approach). However, it should be noted that in an RCT with infrequent outcome rates, a stricter benchmark for denoting strong concerns about attrition than 20% (e.g. 10-15%, or even less) might be more appropriate.3

An educator can always consider demonstrating the calculation used by Guyatt and colleagues to trainees who are evaluating an RCT examining an infrequent outcome. Alternatively, the educator might want to simply point out verbally that, if the loss during follow-up rate is close to the outcome rate, that this circumstance can be seen as especially concerning (since it may result in bias that quite substantially alters a RCT’s findings).

We have also noticed manuscripts in which a less extreme form of these calculations were performed as a sensitivity analysis to assess potential bias from loss during follow-up. For example, one study used the outcome rate seen in the control group and assume this rate of outcomes existed for those participants lost during follow-up in the control group.22 While this assumption is probably more realistic than the “worst case” scenario, this approach ignores the likelihood that the participants lost during follow-up may have a worse prognosis than participants remaining in the trial (even those participants in the control group).

 In the Tutorial, we do not provide a benchmark for what amount of difference in loss during follow-up might be considered “differential.” Benchmarks that might be considered for differential loss during follow-up that could be judged as particularly concerning might be a ≥ 5% difference in attrition between study arms (although in studies with particularly high rates of attrition, such as 30% or higher, this might be relaxed to a ≥7.5% difference in attrition between study arms). However, this is based on our opinion rather than data. (Determining what degree of differential loss during follow-up appears to substantially bias study findings might be a useful topic of a meta-analysis). This 5+% benchmark may sound rather stringent, since many trials may have loss during follow-up rates that differ this much between study arms. However, we suggest it because it is very difficult to envision scenarios in which, once randomization has balanced the two study arms, that different percentages of individuals can be lost from the different study arms without substantially disrupting the initial balance in participant characteristics established by randomization. Furthermore, if participants are stopping their participation in a RCT due to adverse outcomes, this could add to the impact of even relatively modest difference in attrition (i.e., ≥ 5%) on the study findings.

 It should also be noted that methods that estimate the outcomes for individual lost to follow-up, such as last-observation-carried-forward and multiple imputation, may mitigate some of the potential bias that could arise from loss during follow-up. However, since these missing outcomes are estimated, it is impossible to know how well they actually address the potential bias introduced by loss during follow-up. (That is, it is impossible to know how well the estimate matches the outcomes that actually would be observed should these individuals remain in the trial). A detailed discussion of the advantages and disadvantages of last-observation-carried-forward, multiple imputation and other estimation methods is beyond the scope of our Tutorial and the discussion here. We do note in the Tutorial that multiple imputation is generally expected to do a better job reducing bias from loss during follow-up than the last-observation-carried-forward method.20 However, trainees or users who are particularly interested in statistics and or methods for addressing missing data might want to learn more about these, and possibly other, methods for addressing loss during follow-up.

The existence of these methods, although they are not perfect, may be one reason why a 2016 study found that the bias attributable to attrition was more variable than observed from other factors, such as blinding.11

**Outcomes**

One important detail pertaining to multiple comparisons is that US Food and Drug Administration’s (FDA) policy specifies that, if a primary outcome is clearly declared, then that primary analysis does not have to be adjusted for multiple comparisons even when the trial is also examining other outcomes (such as secondary and safety outcomes).23

**Effects**

We wish to note that the purpose of the sample calculations of the Bonferroni adjustment in the Tutorial was not to advocate for the Bonferroni adjustment. Many other approaches for adjusting for multiple comparisons exist,24 and, as indicated in the Tutorial, not all researchers agree that corrections for multiple comparisons are even needed.25, 26 Rather, we discussed the Bonferroni adjustment simply to highlight that a relatively quick calculation exists that can help indicate whether or not a statistically significant association reported in a manuscript that did *not* adjust for multiple comparisons would still be found to be statistically significant if such adjustments were performed. The Bonferroni adjustment is a good choice for this evaluation since, in addition to being a quick calculation, it is widely viewed as conservative27 (that is, potentially over-adjusting significance levels). Because of this property, if an association remains significant even when a Bonferroni adjustment to significance levels is done, it is likely it would be statistically significant if other common methods of adjusting for multiple comparisons are used.

Regarding effect sizes, we found it harder to find universally-accepted benchmarks for small, medium, or large effects for Number-Needed-to-Treat than for Cohen’s *d*. The range given for small effects in the Tutorial is a combination of a threshold of 10 provided by Citrome and colleagues,28 and a manuscript that converted Cohen’s *d* benchmarks for small effects to Number-Needed-to-Treat.29 (Strictly speaking this study found a Cohen’s *d* of 0.2 equaled a Number-Needed-to-Treat of slightly greater than 13, which we rounded to 15 in hopes of making the benchmark more easily remembered).29 In general, the benchmarks for Number-Needed-to-Treat in relation to effect size appear to be slightly less well delineated in the literature than for Cohen’s *d*.

**References**

1. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med. 2007;26(4):734-753.

2. Schober P, Vetter TR. Correct Baseline Comparisons in a Randomized Trial. Anesth Analg. 2019;129(3):639.

3. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet. 2002;359(9308):781-785.

4. Driscoll P, Lecky F. Article 5. An introduction to estimation--2: from z to t. Emerg Med J. 2001;18(1):65-70.

5. Nuesch E, Trelle S, Reichenbach S, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. BMJ. 2010;341:c3515.

6. Dechartres A, Trinquart L, Boutron I, et al. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. BMJ. 2013;346:f2304.

7. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol. 2000;53(11):1119-1129.

8. Beets MW, Weaver RG, Ioannidis JPA, et al. Influence of pilot and small trials in meta-analyses of behavioral interventions: a meta-epidemiological study. Syst Rev. 2023;12(1):21.

9. Hoge EA, Bui E, Mete M, et al. Mindfulness-Based Stress Reduction vs Escitalopram for the Treatment of Adults With Anxiety Disorders: A Randomized Clinical Trial. JAMA Psychiatry. 2023;80(1):13-21.

10. Moustgaard H, Clayton GL, Jones HE, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ. 2020;368:l6802.

11. Page MJ, Higgins JP, Clayton G, et al. Empirical Evidence of Study Design Biases in Randomized Trials: Systematic Review of Meta-Epidemiological Studies. PLoS One. 2016;11(7):e0159267.

12. Dechartres A, Trinquart L, Faber T, et al. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol. 2016;77:24-37.

13. Hrobjartsson A, Emanuelsson F, Skou Thomsen AS, et al. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol. 2014;43(4):1272-1283.

14. Hrobjartsson A, Thomsen AS, Emanuelsson F, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ. 2012;344:e1119.

15. Altman DG, Schulz KF, Moher D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med. 2001;134(8):663-694.

16. Quitkin FM, Rabkin JG, Gerald J, et al. Validity of clinical trials of antidepressants. Am J Psychiatry. 2000;157(3):327-337.

17. Davis MC, Lee J, Horan WP, et al. Effects of single dose intranasal oxytocin on social cognition in schizophrenia. Schizophr Res. 2013;147(2-3):393-397.

18. Martin GL, Trioux T, Gaudry S, et al. Association Between Lack of Blinding and Mortality Results in Critical Care Randomized Controlled Trials: A Meta-Epidemiological Study. Crit Care Med. 2021;49(10):1800-1811.

19. Abraha I, Cherubini A, Cozzolino F, et al. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. BMJ. 2015;350:h2445.

20. Andrade C. Intent-to-Treat (ITT) vs Completer or Per-Protocol Analysis in Randomized Controlled Trials. Indian J Psychol Med. 2022;44(4):416-418.

21. Guyatt G, Rennie D, and the Evidence-Based Medicine Working Group. Users' guides to the medical literature : a manual for evidence-based clinical practice. (Chapter 1B1. Therapy). Chicago, IL: AMA Press 2002.

22. Irwin MR, Carrillo C, Sadeghi N, et al. Prevention of Incident and Recurrent Major Depression in Older Adults With Insomnia: A Randomized Clinical Trial. JAMA Psychiatry. 2022;79(1):33-41.

23. Food and Drug Administration, Multiple Endpoints in Clinical Trials Guidance for Industry Available: https://www.fda.gov/media/162416/download.

24. Vickerstaff V, Omar RZ, Ambler G. Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. BMC Med Res Methodol. 2019;19(1):129.

25. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1(1):43-46.

26. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. 1998;316(7139):1236-1238.

27. VanderWeele TJ, Mathur MB. Some Desirable Properties of the Bonferroni Correction: Is the Bonferroni Correction Really So Bad? Am J Epidemiol. 2019;188(3):617-618.

28. Citrome L. Quantifying clinical relevance. Innov Clin Neurosci. 2014;11(5-6):26-30.

29. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. Biol Psychiatry. 2006;59(11):990-996.