

TABLE 1

Design Standards and Evaluative Rubric for Between Groups Studies

Design Standard	Evidence for a Yes Rating	Evidence for a Maybe Yes Rating	Evidence for a Maybe No Rating	Evidence for a No Rating
A. How would you rate the alignment of the intervention to commonly-held ideas of the intervention or approach?	The intervention or approach fully reflected commonly-held or theoretically derived ideas about what the intervention or approach should be.	At a minimum the intervention or approach at least somewhat reflected commonly-held or theoretically derived ideas about what the intervention or approach should be.	The intervention or approach was described only as member of broader classes (across which significant variation in content can be expected).	The intervention or approach did not reflect commonly-held or theoretically derived ideas about what it should be
B. How would you rate the implementation and replicability of the intervention or approach?	The intervention or approach was sufficiently described at a level which would allow relatively easy and thorough replication by other implementers, and the description of the implementation of intervention was fully consistent with its defined characteristics.	The intervention or approach adequately described to allow replication of the most essential elements by other implementers, and the description of the implementation was largely consistent with it's defined characteristics.	The authors of the study omit important descriptive information concerning the essential elements of the intervention such that its replication would be impossible, OR it is plausible that the implementation of the intervention may well have been inconsistent with it's defined characteristics.	The authors of the study omit important descriptive information concerning the essential elements of the intervention such that its replication would be impossible, AND it is plausible that the implementation of the intervention may well have been inconsistent with it's defined characteristics
C. How would you rate the adequacy with which the outcome measure was defined?	The study provided adequate evidence that the outcome measure was properly defined and appropriate for the context of the study.	Although the study did not present adequate evidence that the outcome measure was properly defined but the measure did appear to be appropriate to the content of the outcome and the context of the study.	The outcome and/or the measure used to assess the outcome were only described conceptually as a member of a broader class of outcomes/measures about which significant variation exists as to their specific content.	It is unclear what the outcome is and how it was measured.
D. How would you rate the adequacy with which participants in the comparison or	Participants were randomly assigned to conditions, and there does not appear to have been	Either randomized assignment was used but there appears to have been serious	Randomized assignment was not used and despite some steps taken to make the groups	It is unlikely or unknown that/if the participants in the groups are comparable

Design Standard	Evidence for a Yes Rating	Evidence for a Maybe Yes Rating	Evidence for a Maybe No Rating	Evidence for a No Rating
alternative treatment group(s) were made comparable to those in the treatment group?	any serious differential attrition within groups or severe attrition across groups.	differential attrition within groups or serious overall attrition across groups, or although random assignment of participants to groups was not used, there does not appear to have been serious attrition problems within or across groups and reasonable attempts were made to make the groups comparable (i.e. matched sampling, use of a covariate, etc.).	comparable, they do not appear to have been adequate.	
E. How would you rate the adequacy with which the study controlled events that happened concurrently with the intervention or approach that might have confused its effect(s)?	Concurrent processes and events that might be alternative explanations to a treatment effect have been ruled out, either explicitly or implicitly.	There were no identified processes or events that could be alternative explanations for a treatment effect, but some alternative explanations cannot be explicitly ruled out either because there was some evidence that alternative explanations might exist, or because no attention was given to ruling out an alternative explanation and it is reasonable to expect that one or more alternative explanations might exist.	There was no “Maybe No” rating in this standard.	Identifiable processes or events that are described to be occurring simultaneously with the treatment or approach may have caused the observed effect
F. How would you rate the adequacy with which the actual sample, setting, outcome(s), and measurement processes reflected	The actual sample generalizes well to the theoretical population and the setting, outcome(s) and measurement processes generalize	Most aspects of the theoretical population and common variations of settings, classes of outcomes, and data collection	Although some important characteristics of the theoretical population and typical settings, outcomes, and data collection processes	The actual sample does not adequately reflect any characteristics of the theoretical population, and the setting, outcomes,

Design Standard	Evidence for a Yes Rating	Evidence for a Maybe Yes Rating	Evidence for a Maybe No Rating	Evidence for a No Rating
the theoretical population and typical norms for settings, outcomes, and measurement processes?	well to common variations in settings, classes of outcome(s), and processes and timing of data collection.	processes and timing are represented in the study.	and timing are represented by the study, many important characteristics are not.	and data collection timing and processes have characteristics that are not within the boundaries of accepted and typical practice.
G. How broadly was the intervention tested statistically across important sub-groups of students, and across substantive variations within the intervention as a whole?	The analyses in the study examined the effect(s) of the intervention across important sub-groups of students AND included separate analyses of key sub-components of the intervention for differential effectiveness on those different sub-groups of students.	Some sub-group analyses were conducted and some estimates were made exploring differential effects of different intervention components.	Some sub-group analyses were conducted or some estimates were made exploring differential effects of different intervention components. However, significant sub-groups were omitted from the analyses, and no separate effects of different intervention components by sub-groups were explored.	Only main effects for the intervention as a whole were reported with no sub-group or intervention component analyses.
H. How thoroughly were the assumptions underlying the statistical analyses for the study reported?	It is clear from the design that the assumption of independence across groups and observations was not violated, and some evidence is provided that other important assumptions underlying the statistics for the study (i.e. homogeneity of variance) were not violated.	It is clear that the assumption of independence across groups and observations was not violated, but other information about assumptions underlying the statistics for the study are not provided.	It appears that the assumption of independence across groups or observations was likely to be met and other information about assumptions underlying the statistics for the study are not provided.	It does not appear from the study's design that the assumption of independence across groups or observations was met.
I. How adequately were the data described, analyzed, and depicted such that effect size for the outcome in this extraction is able to be calculated?	Either the effect size was reported by the authors or they provided data to allow precise calculation of effect sizes.	Sufficient statistical information was reported to allow, at a minimum, an imprecise effect size to be calculated for the outcome of this extraction.	There was no "Maybe No" rating in this standard,	Neither sample sizes nor effect sizes were reported, and insufficient data were provided to allow those effect sizes to be calculated.