**Associate Editor**

(1) The reviewers appear to have different views on the bias of omitted variables in shifting-case analyses. Reviewer 2 (#4) suggests reminding readers about the omitted variable bias in meta-regression. On the other hand, Reviewer 1 (#9) believes that the bias in shifting-case analyses should compare estimates of these coefficients to the true parameter values in the restricted model. There are many potential covariates in a meta-regression, e.g., year of publications, mean age of participants, and duration of intervention. If we always assume that the model with all covariates is correct, all restricted models with fewer covariates are misspecified. I think that this issue needs further discussion.

This is an excellent point. We have added a new subsection 3.3 to discuss this issue.

(2) Another related issue is the substance abuse interventions example, which includes Group 1 Hi-Int. and Group 2 Hi-Int. as two covariates. It gives readers an impression that we should always include both covariates in the analysis. However, if the covariates are year of publications, mean age of participants, and duration of intervention, does it make sense to include them only one at a time? Moreover, the effect sizes in the example are correlated; this may complicate the results because the paper assumes independent effect sizes (see also Reviewer 1, #1, #2, and #3).

In general, regression modelling guidance suggests caution is warranted when interpreting coefficients in the face of known or theorized omitted variables. We take this approach here: omitting known and observed confounders from a regression model results in omitted variable bias. Whether we refer to it as “bias” or simply caveat our findings with the fact that the model does not adjust for important observed confounders, the interpretation is cloudy when we focus on restricted models. Put another way, we always worry about bias from omitting unobserved confounders in meta-regression (see Lipsey, 2003), which means we should absolutely worry about bias from omitting observed confounders.

Finally, we appreciate this point about correlated versus independent effect sizes. We note that our general result can be expressed in terms of matrix transformations, and that it generalizes to a setting with correlated effects. However, we restrict our interpretive examples to simpler settings with independent effects. We have modified the language on page 2 and 3 to make that clearer. In addition, we have emphasized how our findings and assumptions can be expanded to dependent effect size structure (p. 5).

(3) The shifting-case analysis (SCA) seems the same as the pairwise deletion in data analysis. It will be helpful to mention it so that readers from other disciplines find it easier to follow.

We have made note of this in the introduction: CCA is equivalent to listwise deletion, while SCA is pairwise deletion.  
  
(4) Figure 4 suggests that the omitted variable bias and missingness bias are always in opposite directions. Is it true?

This is not necessarily true based on the results presented. To avoid confusion, we have added the following sentence on page 22: “Note that while the omitted variable bias and missingness bias are in opposite directions in this example, this need not be the case in general; both biases could feasibly be in the same direction for other data.”  
  
(5) The current version does not follow the required format specified in the Author Guidelines. For example, the AMA reference style must be used. Please refer to the Author Guidelines for details.

Apologies for this oversight. We have used the AMA bibliography format for this revision.

**Reviewer 1**

1. Section 2. In the substance abuse interventions example, it is mentioned that “Often the same group (typically the control group) in a study was used in multiple contrasts...”. For a control group, how would a covariate indicating high- versus low-intensity intervention defined? It seems that this would be undefined, rather than missing?

We apologize for the confusion, as we did not accurately convey the nature of these data. Since the studies involved adolescents with substance abuse disorders, most researchers opted not to use a no-treatment/passive control arm, over concerns about failing to treat individuals with a clear disorder. Instead, contrasts typically included two alternative treatments. We have clarified this on page 2: “These effect estimates involve contrasts between groups in a study that are subjected to different treatment conditions, denoted in the data as *Group 1* and *Group 2*, so that each treatment effect can be thought of as Group 1 minus Group 2. Typically, researchers avoided no-treatment or placebo conditions in studies over ethical concerns surrounding the failure to treat adolescents with substance abuse disorders.

Thus, contrasts within studies (i.e., effect estimates) tended to focus on a specific treatment of interest to the researcher versus some alternate treatment.”

1. Section 2. Page 3. The discussion contrasting coefficients for X1 in complete-case analysis (Column 1 in Table 2) and the coefficient for X1 in Shifting-case Group 1 (Column 2 in Table 2; and similarly for X2) is misleading. They are not the same parameter and are usually not expected to the same unless under some special cases. They are not comparable. Whether they are close or different does not necessarily mean any estimate is inaccurate. Both can be correct for their corresponding parameter of interest.

Please see our response to comment (9) below.

1. Section 3. Page 4. It would be helpful to make it clearer that the model formulation (1), individual and joint likelihood contributions (2) and (3) correspond to the setting where there are a total of k studies, and each study contributes one Ti. The main technical results seem to focus on this setting (i.e., Ti ’s are independent) while the example in Section 2 corresponds to the setting where Ti ’s can be correlated.

Good suggestion. We have added two caveats on page 4: “This could correspond to a scenario of *k* independent effect estimates presumably from *k* different studies.” And on page 5: “Note that the substance abuse data contains multiple effect estimates per study that are likely correlated. This differs from the model above. However, we can expand this model to account for dependent effect sizes by…”

1. Page 4, line 2 after model (1), typo in “which is true of some effect sizes”?

We have clarified this language: “This assumption is true of some effect size indices and is a very accurate large-sample approximation for others”

1. Section 3.2, page 6, line 49-50, the coefficients βS is not necessarily a subset of the full vector of coefficients β, the full model with all covariates and a restricted model with a subset of covariates can both hold, the restricted model would correspond to the full model with the remaining variables marginalized out. βS and the corresponding subset of β in the full model are not directly comparable.

Please see our response to point (9) below.

1. Section 4, page 9, I fail to understand expression (11). Let T denote an estimator for a target parameter θ, then bias of an estimator is defined as E(T) − θ. I don’t understand the sentence “The bias term δij refers to the bias induced in effect estimate i due to conditioning on missing pattern Rj ”. In the current meta-regression setting, the effect estimate Ti are obtained irrespective of Xi. The potential bias due to missingness in Xi is only relevant when considering the coefficient of Xi in the meta-regression model relating Ti and Xi.

This is a really good point, and something that got us to think through our language more thoroughly. Essentially, delta is the bias induced by looking only at *Ti* for which *Xi* are observed. Particularly if *Xi* are observed more frequently for larger *Ti*, then if we only use those *Ti* to estimate *θi = Xiβ*, we will likely overestimate *θi* (since we would be using the larger *Ti* to do so). We have revised that section to read: “Here, we see that the expectation of *Ti* given *Xi* and *Ri* can be written as the complete-data expectation *Xiβ* (i.e., the regression model) plus a bias term δ*ij*. The bias term δ*ij* refers to the bias induced in the regression model due to conditioning on missingness pattern , which can affect individual components of η.”

1. Section 4, page 9, please define the terms in equation (12). What is mj? what are fmj (·)’s? With p covariates, there will be 2 p missing patterns. How would model (12) work? Are there additional assumptions required to model the missingness mechanisms?

We have amended the paragraph following equation (12) to specify the nature of *fmj* and *mj*. Note that SCA or CCA condition on a given set of missingness patterns, which we define by script , so that contains a subset of the 2p missingness patterns. In that sense, by defining selection into we are defining selection into analysis. Finally, this model is really flexible in terms of structure, however it is only estimable if *mj* < *k*.

1. Section 5.1, page 12, the wi ’s in (18) and (19) are undefined.

Thank you for pointing this out. We now define *wi* = 1/(τ2 + *vi*) on page 5.

1. Section 6. Please see my comment #2. In linear regression models, coefficients in a restricted model (i.e., based on a subset of covariates) are not necessarily a subset of the coefficients of the same variables in the full model with the presence of other covariates. The bias in shifting-case analyses should compare estimates of these coefficients to the true parameter values in the restricted model. More specifically, use linear regression models as a simple example:

E(Y | X1, X2) = α + β1X1 + β2X2 (1)

E(Y | X1) = α ∗ + β ∗ 1X1 (2)

E(Y | X2) = ˜α + β˜ 2X2 (3)

Note that β ∗ 1 in general is not equal to β1, and β˜ 2 is not equal to β2. There are full distributions such that all models (1)-(3) are correct. When we fit models (2) and (3), the target parameters are β ∗ 1 and β˜ 2, not the original β1 and β2. They have different interpretations. The conditions the authors try to identify are those that yield β ∗ 1 = β1, and β˜ 2 = β2. When these conditions are not met, that does not mean βˆ S will be biased.

This is a fair point that caused some serious debate among us. You are correct that the coefficients are different between the full and restricted models. We disagree with how to interpret that difference. We refer to it as *omitted variable bias*, which is consistent with standard regression textbooks and literature on omitting confounders in linear models.

It is possible to conceive of the restricted model coefficients as distinct parameters in and of themselves as you posit. This would inherently specify a different relationship between a covariate and effect size than is specified by the full model. The question is how to interpret this new coefficient and subsequent estimates of it. Regression textbooks and other literature, including Lipsey (2003), suggest that such coefficients be interpreted with caution since they do not account for observed confounders. It is worth noting that by fitting a series of restricted models, there is some concession that each variable could be relevant in explaining between-effect variation, and hence could be a confounder. In other words, while restricted models can be seen as marginalizations of the full model, the meaning of the resulting coefficients is something that the literature on applied regression tends to question.

Our thinking is this: Concern about omitted variables is almost a given in meta-regression. Even the full model with *X*1 and *X*2 is likely not accounting for *something*. If that “something” is not observed or collected, it may be tough to say how much bias there is. It follows then, that the two restricted models with a single covariate also have omitted variable bias. In this case, the restricted models are missing the same “something” as the full model, but they are also missing a covariate we observe (e.g., *X2*), and for which we can quantify omitted variable bias.

We have added a new subsection 3.3 that clarifies these points and notes that βS can indeed have an alternative interpretation.

**Reviewer 2**

1. The manuscript should include additional discussion around the appropriateness/reasonableness of using the log-linear distribution for the underlying selection models. How might the conclusions and recommendations for practice vary under alternate selection models, and under what real-world scenarios might log-linear models be more or less appropriate?

This is a great question. In general, it is impossible to know if a selection model is properly specified, since the precise missingness mechanism is almost never known in practice. However, we feel the use of log-linear models is reasonable here for a few reasons.

1. The link function has only minor effects on results.
2. The general result is based on a very flexible selection model.
3. Logit links are naturally associated with common understandings of the odds of missingness, in fact, the log odds ratio is a common effect size in meta-analysis.

Since analyses condition on *R* ∈ for some collection of missingness patterns , their bias will depend on the distribution of the event {*R* ∈ }. There are certainly alternative link functions we could use in the GLM framework. Approximations for other link functions are messier (e.g., for probit link functions) but wind up depending on the same general factors (i.e., missingness rate, variance, and correlation of missingness with effect size).

Though we restrict our focus to the GLM framework for selection/missingness, our general result relies on a very flexible model that can account for nonlinear relationships (or interaction terms) between *T*, *v*, or *X* and the probability of selection. Thus, our general result encompasses a wide class of possible missingness mechanisms. Finally, most available software that make likelihood-based adjustments for missing covariates in regression models assume a log-linear selection.

That said, we have added to discussion of the log-linear model on page 10 to make these points more clearly.

2. It seems that additional attention may be needed in the discussion section around the finding that bias will be larger in CCA when there are larger values of between-studies heterogeneity. This seems particularly relevant and noteworthy given that meta-regression is often of greatest interest when there is a large amount of between-studies variance (i.e., this is the exact scenario in which many analysts will pursue estimation of meta-regression models).

Great suggestion. We have added a paragraph to the discussion section (p. 23): “Results for both CCA and SCA suggest that bias due to missingness will tend to increase in magnitude as a function of the total variation in the data. This means that if studies have small sample sizes (i.e., *vi* are large) or there is substantial residual between-effect heterogeneity τ2, the bias of a CCA or SCA will be greater. Because meta-regression is used to explain between-effect variation τ2, models capable of explaining much of that variation will have lower bias in CCA and SCA estimates. However, even very modest amounts of residual variation can still result in substantial bias.”

3. Are the results/conclusions reported here sensitive to any other design or analysis features, such as effect size metric, estimator used for tau-squared, measurement level of the covariate with missing data, and marginal/joint distribution(s) of the covariate(s) with missing data?

Another excellent set of questions. These do not assume any specific effect size metric, so long as the effect estimates are normally distributed. This is true for several metrics on which we conduct meta-analyses (e.g., *z*-transformed correlations), and is a very accurate large sample approximation for others. When the assumptions of our underlying model are violated (effect estimates are not normal or their variance is not known), then these results may not be accurate. We have clarified that assumption on page 4: “This assumption is true of some effect size indices and is a very accurate large-sample approximation for others.”

The results here are not specific to any single variance component (τ2) estimator. In fact, the use of the weight matrix in (4) assumes that tau is known. If it is estimated with bias or substantial uncertainty, we expect the properties of estimators to be even worse. We have added language about this on page 5: “Note that the weights involve the true variance component τ2. In practice, τ2 must be estimated by , and the resulting weights used in analyses can be written . For the sake of simplicity, we use *wi* to derive results in this article, and so results do not depend on variance component estimators. Presumably, use of would induce additional variation into analyses.”

The results assume that covariates are measured at the effect estimate level, but make no assumptions about their underlying distribution. As such the general result is somewhat difficult to unpack, and hence we use examples to elucidate the key properties of our general result for bias.

4. Even though omitted variable bias is not a concern in CCA when considering bias that may emerge from dropping missing data, it is probably worth reminding readers that omitted variable bias is always a concern in meta-regression (regardless of the approach used to address missing data for covariates). Without such clarification, the conclusions as reported could be misinterpreted to imply that omitted variable bias is only a concern in SCA.

Great point! We make a note of that in a new subsection 3.3.

5. The application of Cohen’s rules of thumb for interpreting effect size magnitude is not necessary, and potentially misleading to readers; this should probably be dropped entirely. These rules of thumbs are acontextual and were never intended to be used as universal rules of thumb for interpreting effect size magnitude. It seems better to let the results speak for themselves regarding the resulting bias that may emerge under different scenarios, rather than citing these problematic rules of thumb that are not empirically based (nor empirically validated).

We appreciate this comment, however we feel that justifying values used to compute results better contextualizes the findings than simply omitting explanation. That said, we are sensitive to this idea that they are arbitrary conventions. To that end, we have amended language in relevant sections to point out that while we chose values to align with some known conventions, those conventions are not always based on empirical results, and so may be considered arbitrary.