RESPONSES TO REFEREE REPORT NUMBER 1

I thank the reviewer for his/her comments. My responses follow.

1) COMMENT: First, I am not really seeing the message of this paper. Is it that there is a lot of heterogeneity in this data-set? This, however, is well known. In the introduction it is stated that "dispersion and heterogeneity statistics are used to assess the performance of each method". However, the paper does not really assess the estimators' performance, but merely draws general conclusions from one sample. This sample could serve as an illustration of some message, however, as mentioned, the message of presence of heterogeneity already is quite well known.

RESPONSE: I'm not sure if the referee is commenting on Table 1 and Figure 1 or more generally about the body of research that the paper draws on. Given the sample, there are still issues about the degree of heterogeneity and how it is to be handled in the context of meta-analysis. As indicated in the paper (p. 2 and p. 12), the usual approach in meta-economics is to jump to a FES meta-regression with only cursory considerations of other methods. I think that is unwise and the paper makes the case that a broader set of statistics and methods should be presented. That is the main message in the paper, viz.:

“Despite the high degree of heterogeneity that exists in economic data and studies, an overwhelming majority of meta-analyses in economics employ a conditional or fixed-effects size (FES) model to summarize and analyze study results.” (p. 2)

“The choice of FES vs. RES models is crucial as the starting point for a meta-analysis. However, relatively few meta-analyses in economics dwell on this distinction or inform readers that a FES analysis cannot be extended or generalized beyond the collection of primary studies. . . The present paper addresses similarities and differences between the models and presents results for both weighted-means and meta-regressions for alcohol tax pass-through rates.” (p. 3)

As indicated in the introduction, the most common approach in meta-economics is use of a fixed-effect (FES) model and a FES meta-regression model. Illustrative numbers are cited from Nelson and Kennedy for an earlier survey, but there is no reason to think this has changed much. If heterogeneity in primary studies is measurable using various moderators, then one might apply the FES model (although the weighting in question can be misleading, see p. 7). This is shown not to be the case for the present analysis, where the moderators explain little (Tables 6 and 7). Similarly, if there is no desire to generalize beyond the sample of primary studies, then the FES model can be appropriate. This is stated on page 2 and repeated on pages 17-18. If substantial heterogeneity exists – as the referee indicates – and is well-known, then the issue is still the appropriate way to deal with it. That is the motivation for the paper – what results are obtained for FES vs. RES models and does it matter for this sample or more generally? For the sample in question, this is important for policy applications involving calculation of optimal alcohol taxes, which involve different countries, beverages, market structures, time periods, etc. (p. 4). Similarly, benefit-value transfers in economics are often aided by a meta-analysis, which is mentioned at several places in the paper (p. 4, 28), with limitation to FES models in this context.

Two basic methods are compared by use of a standard set of procedures, which go beyond those reported in Nelson and Moran (see below). I cannot name another paper in meta-economics that engages in this comparison (p. 3 briefly indicates this). Rather than just assert a method, the paper deals with the choice of method. It does not deal with choice of sample, although that is discussed in Nelson and Moran. It should be clear, I think, that this not a Monte Carlo study, so similar analyses might support alternative outcomes for choice of method. Here are statements that illustrate the basic message in the paper:
Page 2 – “However, inferences beyond the sample of estimates should not be carried out using a FES model . . . a FES model assumes the between-study variance is zero or nearly so, which implies perfect homogeneity of true effects. Choice of model is therefore an important aspect of any analysis.”

Page 2 – “The objective of this paper to illustrate how results from a RES model can differ from those in a FES model, reflecting unobserved or random heterogeneity in a sample of estimates.”

To the best of my knowledge, no other paper in economics provides a comparable discussion of models and methods or illustrates statistical outcomes for fixed- versus random-effects meta-analysis.

As a response to the referee’s concerns, I would add the following on Page 3: “As demonstrated below there is substantial dispersion of estimates, which suggests use of a RES model. Application of random-effects methods also may be important for public policy purposes if the tax application in question is for a non-sampled population, i.e., a value transfer is required.”

2) COMMENT: Second, heterogeneity is commonly addressed in economics. Virtually all meta-analyses in economics address heterogeneity either by random effects, or by conducting a meta-regression analysis. Therefore, the proposal of random effects is not new, and it is not clear why it is preferable to fixed effects. In this context, it is not clear what the paper is adding to this issue.

RESPONSE: Certainly, most papers in economics employ a meta-regression and most do this in the context of a FES regression with categorical moderators. I disagree that random-effects is used (widely) in economics. See again Nelson and Kennedy. However, I believe the referee’s statement ignores several differences between the models that are addressed in the paper and are not commonly appreciated in the meta-economics literature:

a. “The RES or plural-effects model is more appropriate if study results are representative of a universe of comparable populations and the objective is to generalize in terms of a grand mean and variation around that mean . . . In contrast, a FES model assumes the between-study variance is zero or nearly so, which implies perfect homogeneity of true effects. Choice of model is therefore an important aspect of any analysis.” (p. 2)

b. “Thus, a RES analysis is designed to facilitate unconditional inferences about (non-sampled) studies that are similar but not identical to sampled studies” (p. 6).

c. “Raw weights in the FES model vary from 625.0 for the most precise estimate to 0.36 for the least precise, with a median weight of 14.8. Relative weights vary from 9.44% to 0.005%, with a median of 0.22%. For the RES model, raw weights vary from 17.5 to 0.36, with a median weight of 8.12. Relative weights vary from 2.57% to 0.05%, with a median of 1.19%.” (p. 7, n. 10). I would add “The weights are substantially different.”

d. “Although dispersion statistics are not commonly reported in meta-analyses conducted by economists, it seems likely the estimates in Table 2 is not unrepresentative of economic data generally. Most primary samples in economics reflect a wide range of data and empirical methods, making assumption of a common effect size untenable (Nelson and Kennedy 2009). If inferences are restricted conditionally as in the case of fixed-effects, the problem is partially solved. However, the danger is that analysts or readers generalize results or seek to make inferences that go beyond the conditional population such as required by policy applications.” (p. 9)
e. “Third, the analyst could estimate meta-regressions, including covariates that control for systematic differences. This is the methodology commonly chosen in economics, although it needs to be emphasized that either FES or RES models can be applied in a meta-regression.” (p. 9)

3) COMMENT: Even in the context where random effects seem to be appropriate, a fixed effects estimator can still be preferable. Tom Stanley argues in several papers (e.g. Stanley and Doucouliagos, 2015) that the RE method does not work well if you deviate from the clean context for which it is proposed, i.e. adding, among other things, publication bias and realize that much of heterogeneity is from observed sources. In economics, data are not as clean as in other fields, with a lot of publication bias, and, as also stated in the paper, a lot of heterogeneity. Therefore, this type of discussion - fixed vs random effects - may fit better in other research fields.

RESPONSE: I agree that there are cases where fixed-effects can be preferable, especially if (1) a large portion of sample variation is measurable; (2) there is no desire to generalize beyond the sample of primary studies; and (3) the between-study variance is likely to be imprecisely measured. Many meta-studies in economics include numerous moderating covariates. The paper makes the point on Page 22 (citing Anderson and Kichkha) that this often looks like data mining. I also think the cell sizes are often so small for categorical variables that standard statistical tests are likely to be misleading. The general literature on meta-regressions argues that cell sizes of at least 10 observations are to be desired. However, Pages 17-18 present a more general discussion of alternative MRA models (citing Stanley and Doucouliagos 2012). I won’t repeat the five points made, but here is the general conclusion:

“As a result of these several considerations, other commentators recommend a random-effects model (e.g., Borenstein et al. 2009: 196; Borenstein 2019: 229; Thompson and Higgins 2002), with Knapp-Hartung adjustment of standard errors on RES regression coefficients (Borenstein 2019: 36). Both FES and RES regressions are reported for all alcohol and selectively by beverage. In the case of fixed-effects, a Q-test can be used to determine if regression estimates are consistent with a FES meta-regression model. (p. 18) I would add here “All of the Q-statistics reported reject the FES-regression model.”

I disagree with the referee’s suggestion that random-effects is a better fit for other fields. It seems to me that the type of empirical research carried out by economists involves heterogeneous data and diverse econometric methods that is not easily summarized by categorical variables. Consider, for example, the primary studies in the present analysis that involve different countries, time periods, beverages, econometric methods, etc. It seems unlikely as a maintained hypothesis that these data share a common effect size and contain little random or unmeasured heterogeneity.

As Borenstein (2019: 32) points out, if the universe is narrowly defined then a fixed-effect model can be appropriate, and the potential impact of publication bias is less. However, the common finding in meta-economics is substantial heterogeneity and substantial publication bias. This suggests the universe is not narrowly defined and the fixed-effect model is not appropriate.

4) COMMENT: Third, in a related paper "Effects of Alcohol Taxation on Prices: A Systematic Review and Meta-Analysis of Pass-Through Rates" with J. Moran, the author already conducted a meta-analysis of pass-through rates. The paper here is an extension of that paper. In the end of the introduction, it is stated that "The present study is an extended sensitivity analysis of [Nelson and
Moran (2019)'s\footnote{Moran (2019)'s conclusion}. The differences between the two papers have to be worked out better. What is done in this paper that was not done in the previous one?

RESPONSE: Aside from the general emphasis on “choice of method” in economics, there are several differences in methods applied and/or reported in the present paper. I would deal with this issue by adding a footnote on Page 4 along the following lines:

"In particular, the following additions are included in the present paper: (1) basic statistics (Table 1 and Figure 1); (2) heterogeneity statistics for H and T and all prediction intervals (Tables 2-3); (3) WLS means (Table 2-3) (4) subgroup analysis for published studies (Table 3); (5) PEESE results (Table 4); (5) Lin-Chu test (Table 4); (6) trim-and-fill results (p. 15); (7) all cumulative meta-analysis results and trace plot (Table 5 and Figure 3); and (8) a broader set of meta-regression results including additional moderating variables and predicted means (Table 6 and 7).

5) COMMENT: The paper somehow gives the feeling to be a spin-off containing results that did not fit in the previous paper. The estimated prediction interval is huge and considerably larger than the standard confidence interval which reflects the large degree of heterogeneity. But then there is not much learned about the pass-through rates.

RESPONSE: Nelson and Moran is an extensive narrative review of the primary studies that focused on the methodologies used in the primary studies and primary results with implications for public policy. It attempts to provide guidelines for future research in the tax incidence literature. The meta-analysis was intentionally kept brief since it seemed likely that public finance journals would be lacking in expertise in this area. As an illustration of the shift in emphasis, Nelson and Moran only included general references to the meta-analysis literature (e.g., Borenstein et al. 2019; Stanley and Doucouliagos 2012). The present paper includes 33 new references that pertain to meta-analysis or related issues, such as benefit-transfer.

An emphasis in the present paper is reporting of confidence intervals, rather than point estimates. This is outlined out on Page 4, n. 4. The following recent paper would be added to that footnote in a revised paper (references to McCloskey’s books and papers could also be added, although I think these are already well-known in the area):


Following Romer’s suggestions, it would be appropriate to give more emphasis to the economic interpretations of both confidence and prediction intervals. I point out on Pages 7 and 9 (and page 22) that the prediction interval is consistent with both under- and overshifting of taxes, which is a crucial point. I would add a footnote on Page 9 along the following lines:

As pointed out by Borenstein (2019: 94), meta-analysis is designed to address to two important issues: (1) estimation of a mean effect, where the 95% confidence interval is a measure of precision of observed effects; and (2) estimation of dispersion of true effects, which is the information shown by the prediction interval. The prediction interval only has meaning in the context of the RES model, since the FES model assumes a common population effect. When the prediction interval is judged to be large in economic magnitude, the effect size can vary
importantly across comparable populations. In the present context, this means that both under-and overshifting of taxes are possible.

6) **COMMENT:** One of the proposals of how to solve excess heterogeneity is concentrating on subgroups of estimates. However, as seen in Table 3, the prediction intervals get even larger. So what do we learn here? Another solution is a meta-regression analysis which deals with observed heterogeneity. Again, FES is compared with RES with a similar conclusion as before. Would it be not interesting to calculate the prediction interval of the mean so that to show that it gets smaller?

**RESPONSE:** What we learn is that prediction intervals are still large, but now tailored to the subgroup level. In part, this reflects changes in the between-study variance. Note that the RES meta-regressions have already incorporated the between-study variance in the estimation method. I would add the prediction interval in Tables 6 and 7, with the following results:

Table 6, 95% PI for Pred. mean, rgr (1), 0.482 – 1.341; rgr (2), 0.437 – 1.349; rgr (3), 0.482 – 1.358.

Table 7, 95% PI for Pred. mean, rgr (1), 0.133 – 1.785; rgr (2), 0.088 – 1.658, rgr (4), 0.595 – 1.227, rgr (5), 0.579 – 1.223.

The prediction intervals are still large, especially those for beer. I would also modify some of the discussion on page 22 to reflect this addition; i.e., “... in Tables 2 and 6 the prediction intervals are on average 0.7 to 1.6 and 0.5 to 1.3, respectively.”

7) **COMMENT:** The employed publication selection bias methods are not really used in economics... So it is not really clear what is learned here?

**RESPONSE:** The paper is not a comparison of alternative meta-regression methods. It is a comparison of fixed- versus random-effects methods, including meta-regressions. Following Stanley and Doucouliagos, the papers cited by the referee are based on fixed-effects. I make this point also on Page 4 citing Anderson and Kichkha 2017; and Button 2019.

Note again that there are two reasons why basic effect sizes vary across primary studies: (1) effect sizes vary systematically across studies and the observable differences can be coded as moderator variables by the meta-analyst; and (2) effect sizes vary across studies according to factors that are unobservable and unrelated to other factors in the analysis. A consequence of the first reason is the “data-mining” in meta-analyses that is criticized by Anderson and Kichkha. I more reasonable approach for economic data it seems to me is to incorporate fewer moderating variables and allow for random effects (and systematic narrative reviews).

What other evidence in the paper is offered in support of this proposition? First, the general description of the primary studies (pp. 4-5) and basic sample statistics (p. 8). Second, all Q-statistics (pp. 7, 9, 10- 11, 19-21). Third, the poor fit of the meta-regressions and unimportance of the moderators (pp. 20-21).

8) **COMMENT:** I do not really see the purpose of the final paragraph which discusses one of several available simulation-based comparison of different meta-estimators in presence of publication bias.

**RESPONSE:** I would drop the last paragraph and add the following footnote:
“Reed (2015: 37) points out that simulation exercises cannot fully capture the complexity of data commonly available to meta-analysts, including primary studies and estimates carried out using different econometric methods for different countries, data sets, and time periods. He recommends that meta-analysts should report results for different estimators, including weighted-means that do not correct for publication bias.”