Report:

The submitted paper presents a competent meta-analysis of the relation between minimum wages and employment. I have a couple of comments and suggestions that might improve the manuscript.

The funnel plot in Figure 2 shows “minimum wage coefficients.” Are these coefficients actually comparable? If they don't have the same units, there is no point in plotting them in one funnel graph and using them in one regression. I recommend to focus on the elasticities instead (Figure 1). Both funnel plots display many outliers, and the authors should consider winsorizing these extreme observations (see Havranek et al., 2015b).

The degree of publication bias found in this paper is very small by all accounts. This finding should be stressed more in the introduction and compared to previous meta-analyses on the topic. Do the corresponding regressions presented in the paper use standard errors clustered at the study level? They should.

The authors use the typical inverse-variance-based weights in their analysis. I believe that weights based on the inverse of the number of estimates reported in each study would be more appropriate in this case, and have five major reasons for this claim. First, although multiple Monte Carlo simulations show that inverse-variance weights bring good results in meta-analysis, these simulations do not consider the case when each study reports several estimates of the effect in question, and moreover if the number of estimates per study varies. When weights are not constant across panels, the interpretation of the weighted results with panel data is unclear, which is why some statistical packages (for example, Stata) do not allow the use of such weights. Second, some method variables do not vary within studies (or their within-study variance is very limited). With multiple estimates reported per study the introduction of inverse-variance weighting brings artificial variation to the study-level variables, because they suddenly vary within-studies. Again, it is not clear how to interpret such results, and there have been no Monte Carlo simulations that would help us with inference.

Third, in meta-analysis the reported standard errors are likely to be endogenous to the reported point estimates. Certain method choices (for example, simple OLS versus instrumental variables) influence both the standard errors and the point estimates (see Havranek, 2015). If the influence of the method on the two statistics goes in the same direction, a large coefficient in the funnel asymmetry test may simply reflect this endogeneity instead of any publication or small-sample bias, and vice versa (moreover, as meta-analysis becomes better known in economics, standard errors themselves might become the target of publication selection in order for researchers to increase the weight of their results in meta-analyses). One solution is to use the inverse of the square root of the number of observations as an instrument for the standard error, because this instrument is proportional to the standard error, but not likely to be correlated with method choices (Havranek, 2015). It is unclear how to interpret results of a specification where the employed weights are potentially endogenous to both the response and explanatory variable.

Fourth, inverse-variance weights are highly sensitive to outliers in precision. In most meta-analyses there are a couple of studies that report very small standard errors for no obvious reasons other than idiosyncratic methodology, and very often they also report small point estimates (this issue is connected to the endogeneity problem). The meta-analyst can either omit these studies, which is difficult to justify, winsorize these observations, or include them as they are. The differences between these three approaches increase dramatically when inverse-variance weights are used. Fifth, the weights based on the inverse of the number of observations reported in a study give each study the same importance, which in my opinion is more intuitive than to give each study a weight based on the number of estimates it reports (which is what happens when we use other weights). It would strengthen the results of the paper if the authors could produce a robustness check using the weights I suggest (such as in Havranek et al., 2015a), and discuss the limitations of inverse-variance-weighting with unbalanced panel data. I would also like to see within estimates (regressions including study dummies).

I am also skeptical about the “general-to-specific” approach, which involves sequential t-tests, and would prefer the use of Bayesian model averaging (BMA, see Havranek et al., 2015b). Sequential t-tests are not statistically valid, because each subsequent test does not take into account that the result is conditional on the previous one. BMA, in contrast, can be thought of as an extension of the typical frequentist practice in which different specifications with various control variables are estimated to evaluate the robustness of results. Because most economics meta-analyses have a large number of explanatory variables, BMA is an attractive method for this field, because it helps tackle the obvious model and parameter uncertainty.

A common objection to BMA is the claim that the method is atheoretical, throwing in many potential explanatory variables and using statistical techniques to find the most important ones. The problem is that in meta-analysis we always have a large number of explanatory variables that might (or might not) potentially influence the reported point estimates. For some of them our economic intuition is stronger, for some of them weaker; nevertheless, we want to control for all the major aspects of data, methodology, and publication characteristics. The economic theory rarely helps us decide which of the variables we should omit, and the choice between BMA and OLS with sequential t-tests is not connected to this issue. But I understand that many authors might have strong priors against Bayesian methods, and in my experience the practical differences between BMA and sequential t-testing are small (although that doesn't have to hold in general). An alternative is to exclude a number of insignificant variables jointly using an F-test, and avoid any sequential testing. For example, test whether all variables with p-vales above 0.2 are jointly insignificant, and, if so, exclude them from the model. See Havranek and Irsova (2011) for an application of this approach.

Thank you for submitting your fine paper to Economics e-journal.

References:

Havranek, Tomas, 2015. “Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting,” Journal of the European Economic Association, forthcoming.

Havranek, Tomas & Irsova, Zuzana, 2011. “Estimating vertical spillovers from FDI: Why results vary and what the true effect is,” Journal of International Economics, Elsevier, 85(2), 234-244.

Havranek, Tomas & Horvath, Roman & Irsova, Zuzana & Rusnak, Marek, 2015a. “Cross-Country Heterogeneity in Intertemporal Substitution,” Journal of International Economics, 96(1), 100-118.

Havranek, Tomas & Rusnak, Marek & Sokolova, Anna, 2015b. “Habit Formation in Consumption: A Meta-Analysis,” Czech National Bank WP 3/2015.2