Response to "[106] Meaningless Means #2: The Average Effect of Nudging, by Academics, is 8.7%"

Stefano DellaVigna and Elizabeth Linos

As always, the Data Colada authors provide food for thought in a series of blogs criticizing meta-analyses. In brief, the criticism in the blog is that (i) meta-analyses provide summary measures that aggregate very different types of inventions and thus provide results that are hard to interpret; and (ii) it is difficult to do proper data control in building the meta-analysis data set so invalid results are common. As we write below, these criticisms are largely fair and, to an extent, they do apply to our *Econometrica* paper (partly by design and partly because our data control was not flawless – see below).

As the same time, before we turn to specifics about our paper, we also can find several reasons to defend the use of meta-analyses. Ironically, defending meta-analyses is an odd endeavor for economists who as a discipline have very rarely done or published meta-analyses and would thus be very open to this type of criticism. But we do not agree with this blanket negative view presented by the blog writers and think that meta-analyses, under some conditions, can be really quite valuable. We identify three main grounds.

First, meta-analyses represent a picture of what researchers read in the literature and, to an extent, believe, *warts and all*. If we are interested in what beliefs researchers will have on a particular topic, the existing literature is a necessary starting point. This is not to say that those beliefs will be *correct*. In our case, we explicitly asked researchers to make predictions about the findings of a meta-analysis of nudges to capture the beliefs of researchers. Those predictions line up remarkably with the findings of our meta-analysis of published nudge experiments: there is an expectations of large effect sizes with nudges. Our paper shows that such view is exaggerated and problematic: nudges should be expected on average to yield modest effect sizes, as we show with an analysis of a second sample – our main sample – of Nudge Unit interventions with governments. These exaggerated beliefs have important implications for how new studies are designed both in terms of power calculations and in terms of what intervention is used. As such, documenting what a field "believes" (including published, peer-reviewed studies with improbably large findings) is an important first step in improving any misperceptions.

Second, meta-analysis in general should be done on a relatively narrow topic, comparing studies that are similar as much as possible with respect to intervention and effect size. For example, studying the impact of active labor market policies on the probability of employment or earnings, as Card, Kluve, and Weber (2018) does, strikes us as an effective application. On a very different topic, a meta-analysis of dictator games, as in Engel (2011), also can effectively aggregate across comparable studies. These types of meta-analyses directly address problem (i) outlined in the blog and indeed can be very informative. For example, Andrews and Kasy (2020) use a narrow meta-analysis of minimum wage policies on employment to show that, surprisingly, there is no evidence of publication bias in that sample.

Third, meta-analyses are useful to examine the extent of publication bias using one of the several estimators that identify, and correct for, such bias. The authors of the blog counter that different estimators can imply different corrections. While that is true, this seems to us a case where we risk letting the perfect be the enemy of the good. Estimators that compare the distribution of effect sizes to the right and the left of the significance threshold provide valuable evidence and typically correct in the

same direction (including in our case). Of course, for this to be valuable, the data quality issues in a meta-analysis should not be too serious (more on this below).

In light of these points, it is a legitimate question why in our *Econometrica* paper we did a meta-analysis of published papers on nudges where surely the range of studies and outcomes studied is very large and thus it is clearly vulnerable to the criticisms raised in the blog. The answer is that this meta-analysis is not the key contribution of our paper which is a meta-analysis of two administrative data sets of interventions run by two "Nudge Units" working with US cities and with the federal government. These main samples have largely comparable interventions, at the minimum in the sense that those are all interventions designed in cooperation with governments to affect a pre-specified outcome of policy relevance. Almost all of the interventions are changes in communication from a government to citizens; further, while the outcome variables targeted vary, they all represent outcomes of policy interest to the cities or federal governments. Thus, we do think that that second sample is significantly less vulnerable to the criticism in the post. Consistent with this more limited variation, indeed, the variation in treatment effect sizes is a lot smaller in these studies than in other samples, exactly because they are more similar in modality and context.

After completing an analysis that gave us the average treatment effect of a "government nudge" in this context, though, we wondered how to contextualize our finding of, on average, a 1.5 percentage point impact of Nudge Unit interventions. How would this finding compare to the approximate effect size in the broader literature? (as in the first goal above). Rather than do a separate meta-analysis that could be accused of cherry-picking, we decided to use existing meta-analyses on nudges; the ones we found, especially Hummel and Maedche, had a broad range, admittedly. We still think that it is a useful comparison for the first goal outlined above – to illustrate the kind of papers readers of the literature see about nudges – and it also enables useful, if imperfect, inferences with respect to publication bias, the third point above. We continue to believe that contextualizing our finding on the average effect of a nudge in government communications is useful by noting how it compares to other academic literature (publication bias included).

A final criticism from the blog is about the data control. Surely if there is poor data control, meta-analyses are meaningless. We agree! Indeed, we spent many hours, ourselves as authors, as well as a team of research assistants, pouring over the 100 studies in Hummel and Maedche, and taking out a majority of them because, for example, they were not experiments, or they did not fit the sample criteria (such as a 0-1 dependent variable). By no means did we do this in a cavalier way. In doing so, we also recognized the difficulty of comparing across these studies and of doing excellent quality control over such large range of studies. In particular, while some articles provided transparent evidence from a clear RCT design, others had more limited detail on the experimental design, or results presented in forms that made it hard to compare to other settings. Still, we wanted to present results from an existing meta-analysis to capture the literature, and we believed that we had done a good enough job of cleaning the studies.

Thus, when we read that among the largest 7 effect sizes in our sample, three studies are invalid we were *horrified*. Did we really miss a majority of the data checks despite what we thought were our best efforts? We pride ourselves on being data oriented, perhaps even data obsessive, and we did not take the data problems lightly. We thus decided that we wanted to get a further check of data quality, going beyond the 7 effect sizes checked in the blog, which is a small sample. We decided to expand the data

check to all the top half of effect sizes, 37 effect sizes out of 74. We checked each of the 15 additional papers in this sample. We find that 14 out of 15 are valid experiments and, we believe, we coded the outcomes correctly. We did identify one study where the variation across experimental conditions is varied once per week over 6 weeks, providing insufficient grounds for inference. That is one example we would rather have taken out as well from the sample. To us, this says that, yes, there were a couple of questionable cases, which we regret including in hindsight, but the large majority of the studies we meta-analyze are, yes, highly heterogeneous, but experimentally valid. And once again, we stress that these studies are not the ones we would have picked as most credible examples of nudges, as that was *not* the goal of our exercise. We wanted to present the effect size of the average nudge in the literature, without picking other studies ourselves, which could have been arbitrary.

Finally, turning to the cases outlined in the blog, we agree in retrospect on the first case, the study on bananas; we missed that the comparison group was from a previous week. (A second study in that same paper does not have that issue.) The second case is less obvious to us. We absolutely agree that retracted papers should not be included in meta-analyses, so we explicitly checked that this paper was not retracted at the time of publication, and it was not. We decide that it was right to include it, again to avoid "cherry-picking" findings we believed. As for the third case, on JAMA pediatrics, coding the results was tricky because in general we tried to transcribe results from the simplest specifications to keep things transparent. In this case, though, with just 12 units of randomization and an imbalance ex ante (which we did not pay enough attention to) it is correct that the diff-in-diff is a better estimate. Finally, we identified all the nudge treatments, but it is a fair point that we should have compared the Nudge+Chef treatment to the Chef treatment, not the baseline. Certainly, this does go to show that it is hard to do good perfect quality control, and that it will not be perfect even with high effort (which we are convinced we put). It also shows that errors are indeed most likely to contribute extreme results.

Ultimately, how much would this change our results? Given that we recoded all the top half of the studies, we re-run our specification taking out cases (1) and (3) as well as case (2), even though we find that less obvious. We find that the average treatment effect for the published nudges is 6.4 pp., which is still over 4 times the average effect of the Nudge Units sample.

Taken together, the fundamental findings of our study holds: the academic literature does include publication bias; scholars and practitioners alike internalize this publication bias to assume that a light-touch *government* nudge will be more effective than it is in reality. Conversely, when the full file drawer of government nudges are analyzed, effects are more modest but still significantly positive.

On the topic of meta-analyses, we are thankful for this blog contribution which illustrates two very important, and valid, points. We will work hard to keep them into account moving forward, as we hope others will do too. As we wrote in this note, we do think there are also important benefits to meta-analyses done the appropriate way, which should be weighed against those risks.