Jay P. Greene's Response to Supplement Submitted by C. Kirabo Jackson

This responds to issues raised by the Supplement submitted by C. Kirabo Jackson on June 22, 2020 and amended on July 9, 2020.

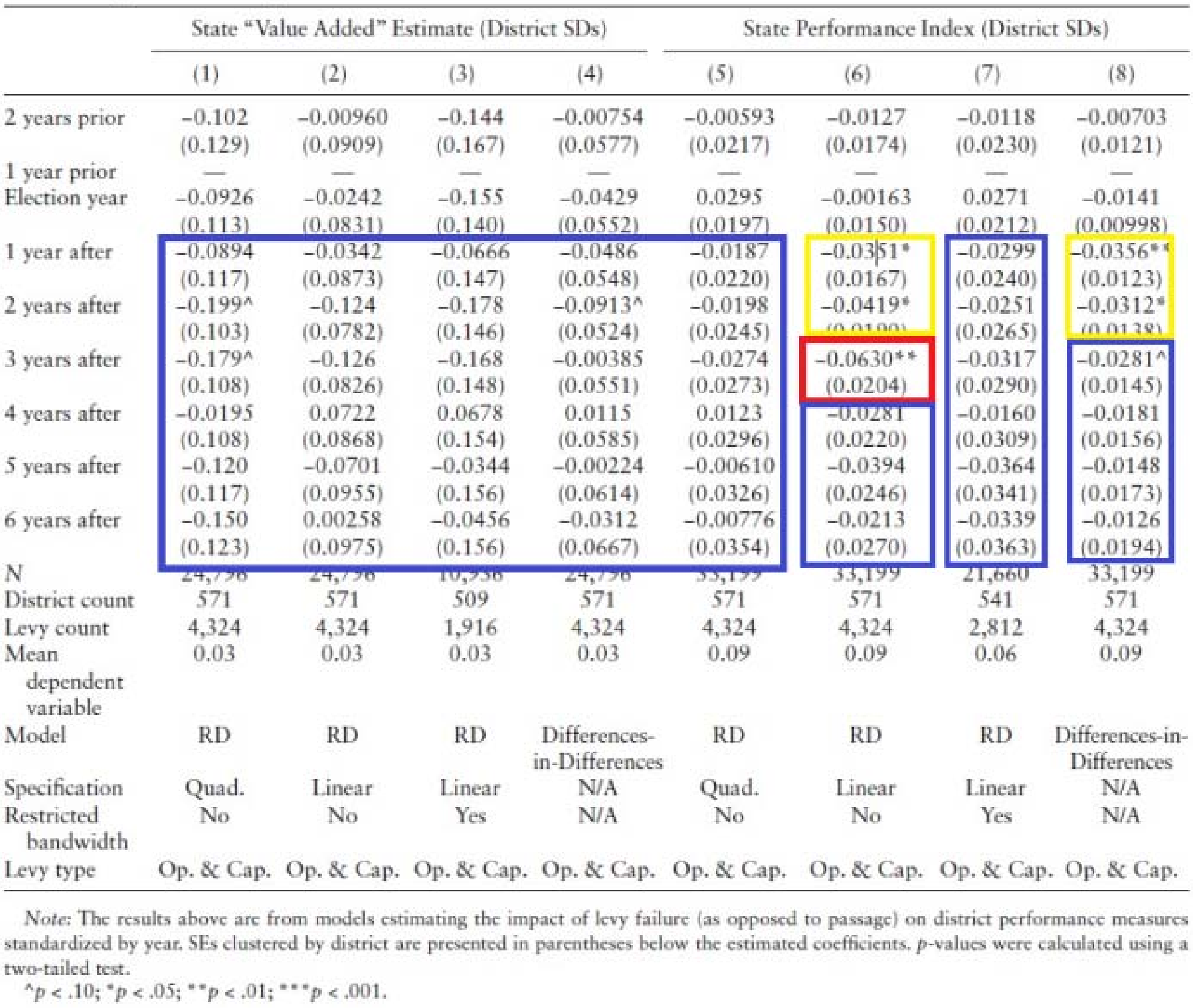
Problems with Calculations in Jackson's Meta-Analysis

1) Subsequent to filing his Supplement disclosing that he relied upon calculations for classifying the results in his Expert Report, Dr. Jackson provided the "SchoolSpendingPapersACLUStatic" and "CapitalProjectsYrXYr" worksheets he used for those calculations. Those worksheets reveal a number of problems with his analysis:
    a. **Missing Studies** – Two studies cannot be found in the worksheets, Biasi (2019) and Card and Payne (2002). Given Jackson's emphasis in the Supplement that his Report classified the results of studies based on his calculations, it is unclear how he could have classified those two studies if they are absent from his worksheets.
    b. **Arbitrary and Biased Selection of Findings Within Studies** – In my Expert Report, I noted that most studies that Jackson reviews contain multiple analyses, with different outcomes, different model specifications, and for different time periods. To handle the multiplicity of results within each study, meta-analyses must either consolidate results within a study, or have a decision-rule that is clearly articulated and consistently applied by which the meta-analysis selects which finding to treat as the main result. In Exhibit 5 of my Expert Report, I illustrated this problem by showing that Kogan, et al (2017) presented 48 different results in a single table. A key question is whether Jackson had a decision-rule for selecting one of these 48 findings as the main finding to characterize Kogan, et al.

       The worksheets with his calculations reveal that rather than being guided by a clearly stated and consistently applied decision-rule, Jackson chose to focus on certain findings within each study in an arbitrary way that biases his overall assessment of the research literature. In the case of Kogan, et al, the worksheets reveal that of the 48 results Jackson could have selected, he chose the statistically significant finding with the largest effect. I have highlighted the result Jackson chose with a red box in the image below. He could have chosen one of the four other statistically significant results, highlighted in yellow boxes, which show smaller benefits from additional spending. He also could have selected any one of the 43 other results, highlighted in blue boxes, that show no statistically significant relationship between additional spending and student outcomes.

Why did Jackson choose the one result in the red box? It does not utilize the more rigorous, value-added measure of student outcomes that control for past student achievement. It does not use the most rigorous model specification that restricts the bandwidth of cases to those close to the threshold of passing a levy increase, which might plausibly be thought of as approximating a random assignment experiment. It does not consider longer term outcomes, as Jackson prefers to do with capital spending analyses. It is difficult to explain the selection of the one result in the red box other than that it is most favorable to the claim that additional spending improves student outcomes.

Table 6. Impact ofTax Levy Failure on Student Achievement



	State "Value Added" Estimate (District SDs)				State Performance Index (District SDs)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2 years prior	−0.102	−0.00960	−0.144	−0.00754	−0.00593	−0.0127	−0.0118	−0.00703
	(0.129)	(0.0909)	(0.167)	(0.0577)	(0.0217)	(0.0174)	(0.0230)	(0.0121)
1 year prior	—	—	—	—	—	—	—	—
Election year	−0.0926	−0.0242	−0.155	−0.0429	0.0295	−0.00163	0.0271	−0.0141
	(0.113)	(0.0831)	(0.140)	(0.0552)	(0.0197)	(0.0150)	(0.0212)	(0.00998)
1 year after	−0.0894	−0.0342	−0.0666	−0.0486	−0.0187	−0.0351*	−0.0299	−0.0356**
	(0.117)	(0.0873)	(0.147)	(0.0548)	(0.0220)	(0.0167)	(0.0240)	(0.0123)
2 years after	−0.199^	−0.124	−0.178	−0.0913^	−0.0198	−0.0419*	−0.0251	−0.0312*
	(0.103)	(0.0782)	(0.146)	(0.0524)	(0.0245)	(0.0190)	(0.0265)	(0.0138)
3 years after	−0.179^	−0.126	−0.168	−0.00385	−0.0274	−0.0630**	−0.0317	−0.0281^
	(0.108)	(0.0826)	(0.148)	(0.0551)	(0.0273)	(0.0204)	(0.0290)	(0.0145)
4 years after	−0.0195	0.0722	0.0678	0.0115	0.0123	−0.0281	−0.0160	−0.0181
	(0.108)	(0.0868)	(0.154)	(0.0585)	(0.0296)	(0.0220)	(0.0309)	(0.0156)
5 years after	−0.120	−0.0701	−0.0344	−0.00224	−0.00610	−0.0394	−0.0364	−0.0148
	(0.117)	(0.0955)	(0.156)	(0.0614)	(0.0326)	(0.0246)	(0.0341)	(0.0173)
6 years after	−0.150	0.00258	−0.0456	−0.0312	−0.00776	−0.0213	−0.0339	−0.0126
	(0.123)	(0.0975)	(0.156)	(0.0667)	(0.0354)	(0.0270)	(0.0363)	(0.0194)
N	24,796	24,796	10,936	24,796	33,199	33,199	21,660	33,199
District count	571	571	509	571	571	571	541	571
Levy count	4,324	4,324	1,916	4,324	4,324	4,324	2,812	4,324
Mean dependent variable	0.03	0.03	0.03	0.03	0.09	0.09	0.06	0.09
Model	RD	RD	RD	Differences-in-Differences	RD	RD	RD	Differences-in-Differences
Specification	Quad.	Linear	Linear	N/A	Quad.	Linear	Linear	N/A
Restricted bandwidth	No	No	Yes	N/A	No	No	Yes	N/A
Levy type	Op. & Cap.	Op. & Cap.	Op. & Cap.	Op. & Cap.	Op. & Cap.	Op. & Cap.	Op. & Cap.	Op. & Cap.

Note: The results above are from models estimating the impact of levy failure (as opposed to passage) on district performance measures standardized by year. SEs clustered by district are presented in parentheses below the estimated coefficients. *p*-values were calculated using a two-tailed test.

^*p* < .10; **p* < .05; ***p* < .01; ****p* < .001.

Similarly, the worksheets reveal that the only way Jackson is able to classify Weinstein, et al (2009) as having positive but not statistically significant effects is to ignore all of the negative results in Table 7 on the effect of Title I on student test scores. (See Exhibit 6 in Greene's Expert Report) Instead, Jackson chooses to focus on the graduation rate result in Table 8. In other studies, Jackson reports test score and attainment results separately. But when analyzing Weinstein, et al, the negative test score results do not appear in his worksheets.

Ignoring a set of results is also required for Jackson to classify Goncalves (2015) as having positive but insignificant outcomes, as opposed to negative and significant results. As Table 4 in Goncalves shows, raising taxes for school construction produces several years of reduced student achievement without improving test scores after the projects are completed. (See Exhibit 7 in Greene's Expert Report). A note in one of Jackson's worksheets reveals that they are aware of this issue: "take estimates of completion exposure for post, how to use construction exposure?" The way he chose to "use construction exposure" is to ignore it because those negative effects do not appear in his calculations.

Jackson also makes an unusual choice in selecting which results to highlight in Abbot, et al (2019). The authors prefer the results presented in Table 8: "Because the more conservative estimates in Table 8 should account for potential changes in student composition, we consider these our preferred estimates of spending's achievement and attainment effects." (p. 9) If Jackson had focused on the results in Table 8 for outcomes in the first five years, he would have reported that additional spending did not have a statistically significant effect on student test scores or attainment. Instead, the worksheet calculations reveal that Jackson focuses on the results for the first five years in Table 6, which do not control for changes in student composition of schools following passage of bond referenda. Only by focusing on those less rigorous findings in Table 6 does he report positive and statistically significant results.

I note that in another meta-analysis that Jackson and his research assistant are conducting contemporaneously, they classify both Abbot, et al and Kogan, et al as not having statistically significant results. (See Jackson, Wigger, and Xiong, 2018, Figure 5 presented as Exhibit 4 in Greene's Expert Report.)

c) **Non-Reproducibility** – For several studies it is not possible to identify where Jackson derives the numbers for his calculations from those studies, making it virtually impossible to reproduce his results. For example, for Baron (2019), the worksheet shows that the effect "0.055" was obtained from "Figure 13 panel c % advanced or proficient." The number, 0.055, does not appear in that figure and a search for it within the document shows no results. I had similar difficulty in finding the numbers on which Jackson relies for his calculations in the locations listed as their source for Kreisman and Steinberg (2019), Conlin and Thompson (2017), and Cellini, Ferreira, & Rothstein (2010).

Discrepancies in Jackson's Classification of Study Findings

2) Dr. Jackson claims that the discrepancies between how he classifies the results of studies in his 2018 NBER paper and his 2020 Expert Report "are by design, and by no means an error." He continues, "Specifically, there are 'apparent' differences because the 2018 summary paper indicates the conclusions of the papers as they were reported by the authors. In contrast, the expert report performs a formal meta-analysis and is based on the results of calculations I made using the numbers reported in the papers." (Jackson's Supplement, p. 1)

Support for this explanation for classifying the results of the same studies differently cannot be found in either the 2018 paper or the 2020 report. In neither document does Jackson articulate the standards by which studies are coded as having overall positive or negative results or whether those overall results are statistically significant or not, nor is there any reason to believe that those standards are supposed to be different in these two reviews Jackson conducted.

In addition, the explanation for classifying the results of the same studies differently that Jackson offers in the Supplement is not consistent with how Jackson actually coded results in either instance. For the 2018 review, Jackson says he relied upon "the conclusions of the papers as they were reported by the authors" to classify results. Yet, the way in which Jackson classifies studies in 2018 is frequently at odds with how the authors of studies describe their own results.

For example, Jackson classifies Weinstein, et al (2009) in both the 2018 and 2020 reviews as positive but not statistically significant. Yet Weinstein, et al describe their own results as negative: "As to the effect of Title I, although none of the coefficients on the Title I eligibility dummy is significant at the 5% level, a number are negative and significant at the 10% level, indicating that there may be a slight negative local average treatment effect in the range of -.03 to -.04 standard deviations. While these findings are disturbing, there may be several reasons why we see these negative impacts." (p. 19)

Similarly, Jackson classifies Van der Klaauw (2008) as having positive but statistically insignificant results in his 2018 review (he excludes this study from the 2020 review). But Van der Klaauw describes his own results as negative: "The estimates indicate that Title I has been ineffective at raising student performance, and in fact appears to have had adverse effects during the 1993 and 1997 school years. Less evidence of adverse effects is found for 2001." (p. 732)

Jackson classifies Goncalves (2015) in both the 2018 and 2020 reviews as having positive but statistically insignificant results. But this is not how the author describes his own results. He says, "I find strong evidence of a negative effect during the construction period. Construction leads to drops of 2.2% of students proficient in math and 1.6% proficient in reading after 4 years of exposure, off a base of 80% proficiency rates. I do not find any significant evidence for positive effects from project completion." (pp. 13-14)

These examples indicate that whatever standard Jackson used for classifying the results of research in his 2018 review, he could not have been simply relying upon how authors describe their own work.

Jackson also states that the 2020 expert report shows different results because it is "a formal meta-analysis and is based on the results of calculations I made using the numbers reported in the papers." The problem with this claim is that Jackson has co-authored a third review of the literature, found in Jackson, Wigger, and Xiong (2018), that is also described as a meta-analysis and yet classifies the results of studies differently. Specifically, Figure 5 of Jackson, Wigger, and Xiong (2018) presents findings from "an ongoing meta-analysis of school spending studies being conducted by Kirabo Jackson and Claire Mackevicius" that characterize results differently from the 2020 report for six out of the 11 studies it considers.[1] This is especially puzzling because Mackevicius is Jackson's research assistant for the 2020 report and the date on Jackson, Wigger, and Xiong (2018) is February 27, 2020 (despite also being listed as a 2018 working paper), which is roughly contemporaneous with the March 12, 2020 expert report presenting different results.

Jackson has produced four different summaries of the research literature on the relationship between spending and student outcomes over the last two years, none of which has the same results as any of the others. He has the 2018 NBER paper, Jackson, Wigger, and Xiong (which was updated most recently on February 27, 2020), the expert report on March 12, 2020, and the report amended on July 9, 2020. The March 12 report differs from the 2018 review in classifying nine of 28 studies, differs from Jackson, Wigger, and Xiong in classifying 6 of 11 studies, and differs from the amended version in classifying 2 of 33 studies. That is a total of 17 discrepancies across these four reviews.

Rather than satisfactorily explaining differences in how Jackson has classified the results of study findings, his Supplement highlights additional discrepancies and therefore provides further support for the opinion I expressed in my expert report that "Jackson's

[1] See Exhibit 8 in Jay P. Greene's Expert Report.

Report is so riddled with errors, inconsistencies, and ambiguities that it is not a reliable summary of that evidence." (p. 4)

Formal Tests Confirm Asymmetry in Jackson's Set of Studies, Indicating Bias

3) Dr. Jackson's discussion of publication bias in his Supplement fails to address the issue of whether his list of 33 studies is a complete and unbiased set of studies to consider when assessing the relationship between additional school spending and student outcomes. To help demonstrate that Jackson was missing studies, I produced a histogram in my Expert Report that showed the bulk of studies that Jackson considers have results that are barely positive – with effect sizes between 0 and .05.[2] If his set of studies were complete and unbiased, we should expect the distribution of results to be symmetrically distributed around this modal, or most common, cluster of results between 0 and .05. Instead, he claims that there are no studies with findings that are lower than the modal result, which would be negative outcomes, and 18 studies with findings that are higher than the modal result. This pattern is clearly not symmetrical and raises serious concerns that Jackson's set of studies is incomplete and biased.

   Jackson's response to this evidence is to note that I "did not present a formal test" of whether the pattern of results is asymmetrical. (p.3) The purpose of a graphical display, like the histogram I presented, is to illustrate results in a transparent way that would be accessible to non-technical readers. The histogram I presented shows what looks like half of a bell-shaped curve, with the negative half of the bell missing. The asymmetry of that visual display is obvious. While a formal test is not necessary, doing something like the Fisher Exact Test, which Dr. Jackson also uses, confirms the strong asymmetry displayed in the histogram. If results are supposed to be symmetrical, the odds that 18 independent studies would have outcomes that were greater than the modal result and none would have results that were lesser would be the same as flipping a coin 18 times and getting 18 heads in a row, or 1/(2^18). Those odds would be 1 in 262,144.

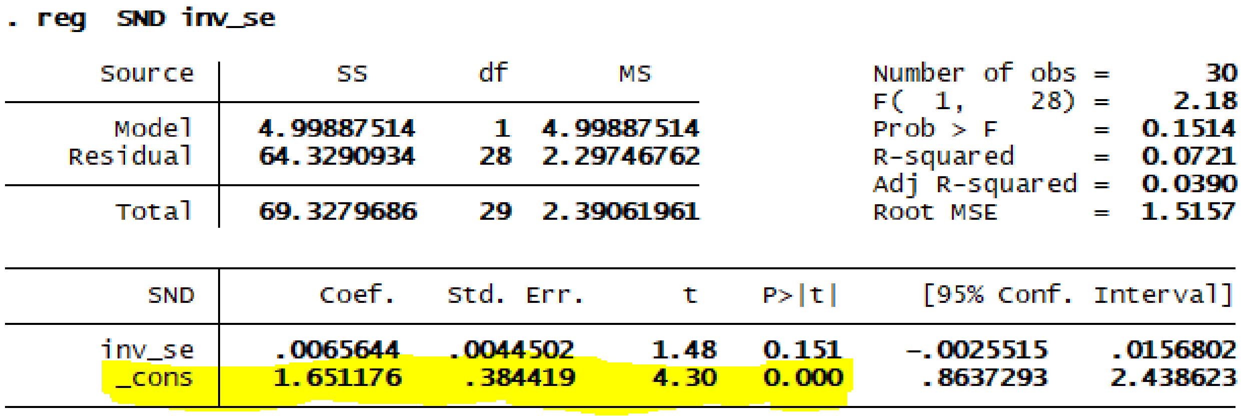
   Jackson does not address the clear asymmetry displayed in the histogram I presented. Instead, he claims to have conducted a formal test for asymmetry in a funnel plot, which I did not present, but which I did reference in making the argument that the results of studies in meta-analyses should be symmetrically distributed. A funnel plot would be like the histogram I presented, but it examines whether effects are symmetrical when controlling for the precision of studies. Jackson says he conducted a formal test and concludes in his amended Supplement that "formal statistical tests fail to show consistent evidence of publication bias." (p. 3)

[2] See Exhibits 9 and 10 in Jay P. Greene's Expert Report.

After it was requested, Jackson provided a data set, “JacksonFunnelReplication.dta,” that he used for this analysis. That data set reveals a number of problems. First, the data set only contains information on 25 of the 33 studies from Jackson’s list. It is not clear why eight studies are missing or whether their absence would alter any results. Second, five of the studies have two sets of “overall” results. It is unclear why five studies should appear twice; perhaps one set presents results for educational attainment and the other for test score achievement, but once again Jackson’s analyses lack clear explanation and transparency. Third, a formal test using the data Dr. Jackson provided contradicts his claim and clearly shows that the results of studies are asymmetrically distributed. Evidence of asymmetry is consistent with publication bias or other problems with the completeness or accuracy of Jackson’s set of studies.

To determine whether the results in Jackson’s data set are symmetrically distributed, I used the Egger Test because it is the most prominent test for asymmetry in funnel plots and the only formal test in the work that both Jackson and I cite. Egger, et al describe the test: “We used a linear regression approach to measure funnel plot asymmetry on the natural logarithm scale of the odds ratio…. The standard normal deviate (SND), defined as the odds ratio divided by its standard error, is regressed against the estimate's precision, the latter being defined as the inverse of the standard error (regression equation: SND= *a*+ *b*xprecision)…. The points from a homogeneous set of trials, not distorted by selection bias, will thus scatter about a line that runs through the origin at standard normal deviate zero (*a*=0), with the slope *b* indicating the size and direction of effect. This situation corresponds to a symmetrical funnel plot. If there is asymmetry, with smaller studies showing effects that differ systematically from larger studies, the regression line will not run through the origin. The intercept *a* provides a measure of asymmetry—the larger its deviation from zero the more pronounced the asymmetry.” (Egger, et al, 1997)

When I run the regression Egger, et al describe, the coefficient for the intercept, or constant, is significantly different from zero. (See printout of result below with the relevant result highlighted.) That means that the regression line does not run through the origin, and there is strong evidence of asymmetry in the distribution of results. Jackson is mistaken when he claims that a formal test would fail to show evidence of asymmetry.



```
. reg  SND inv_se

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(  1,     28)  =      2.18
       Model |  4.99887514         1  4.99887514   Prob > F        =    0.1514
    Residual |  64.3290934        28  2.29746762   R-squared       =    0.0721
-------------+----------------------------------   Adj R-squared   =    0.0390
       Total |  69.3279686        29  2.39061961   Root MSE        =    1.5157

------------------------------------------------------------------------------
         SND |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      inv_se |   .0065644   .0044502     1.48   0.151    -.0025515    .0156802
       _cons |   1.651176    .384419     4.30   0.000     .8637293    2.438623
------------------------------------------------------------------------------
```

The asymmetry of the results in Jackson's set of 33 studies is obvious from the histogram, supported by a formal test for asymmetry of that histogram, and supported by a formal test of a funnel plot that adjusts for study precision. But Jackson claims that even if there were evidence of asymmetry, that "does not in and of itself indicate publication bias." (p. 3) He then cites Sterne, et al (2011) to support this claim. It is true that Sterne, et al describe other possible explanations for asymmetrical results, but none of the other possible explanations strengthen confidence in relying upon a meta-analysis of an asymmetrical set of studies, like Jackson's, for making important public policy decisions. The other explanations for asymmetry that Sterne, et al offer in addition to publication bias are: "Selective outcome reporting," "Selective analysis reporting," "Poor methodological design," "Inadequate analysis," "Fraud," "True heterogeneity," "Artefactual," and "Chance." (Box 1)

Every study of a policy intervention that requires additional funding could be seen as evidence of whether additional resources are helpful or not. But, even if we accepted the constraints of Jackson's preferred methodologies, Jackson only selects a few dozen of those studies to consider, while excluding scores of others. The analysis of asymmetry provided here strongly supports the conclusion that the set of studies Jackson considers is arbitrarily limited and likely to be biased.

Jay P Greene

Jay P. Greene, Ph.D.

Dated July 17, 2020