Reply Expert Report of Clement Kirabo Jackson

The old evidence should not be trusted.

According to common statistical standards, observational studies (which comprise all studies examined by Dr. Rivkin (and his collaborator Eric Hanushek)) are not deemed credible. The standard for what constitutes good policy evidence has changed drastically between the 1970s and the present. The studies referred to in Dr. Rivkin's report are old studies that use methods that are not deemed credible. It is well known that quasi-experimental studies are credible for making causal policy claims, while observational studies are not. This accepted wisdom is reflected in the Standards Handbook of the What Works Clearinghouse (WCC).[1] The WWC is an initiative of the U.S. Department of Education's Institute of Education Sciences (IES), which was established under the Education Sciences Reform Act of 2002. WWC is "part of IES's strategy to use rigorous and relevant research, evaluation, and statistics to improve our nation's education system. The mission of the WWC is to be a central and trusted source of scientific evidence for what works in education." *Id.*, p. 2. To accomplish that mission, the WWC examines studies that have met eligibility screens. Studies using randomized controlled trials, quasi-experimental design, regression discontinuity design, and single-case design satisfy the eligibility screens. Observational studies do not. *Id*

This makes clear that my decision not to evaluate older studies that use objectively inferior research designs (for making causal claims) is not an idiosyncratic choice, but is consistent with accepted good practice and what that the United States Department of Education considers a good standard of evidence. Importantly, all of the studies included in the meta-analysis in my report meet the standard of inclusion.

To be clear, no study is perfect. But there are some forms of analysis that are known to be less credible and reliable than others. The studies I evaluate all use a design considered to be reliable, and the studies considered in the Rivkin report use a method not considered to be reliable. The methods those studies use are so unreliable that the U.S. Department of Education does not consider it to provide policy-relevant evidence.

[1] https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

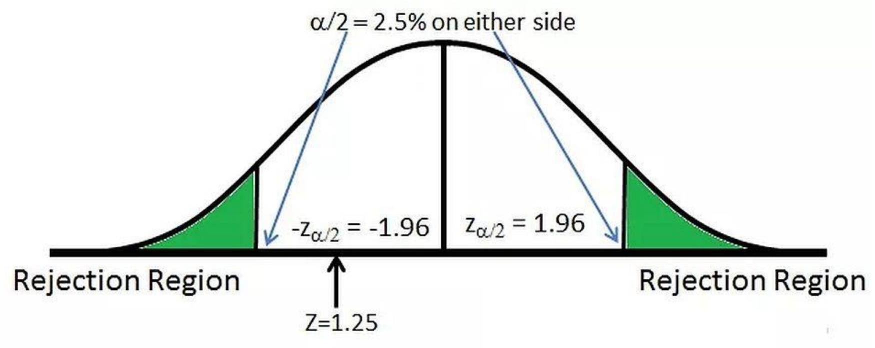
The evidence from the old literature (which has technical problems) does not support Dr. Rivkin's conclusion that there is no systematic "relationship between the quality of schooling as measured by the school's contribution or value added to achievement, educational attainment or future earnings ... and resources." (Rivkin report, p. 4)

Putting aside the fact that the older literature referred to in the Rivkin report should not be trusted because it is based on studies that use unreliable methodologies, his conclusion is based on a statistical error.

To support his conclusion, Dr. Rivkin refers to studies identified by Dr. Hanushek and co-authors, asserting that "only 27 percent of the estimates of the relationship between achievement and per-student expenditure are positive and statistically significant." He suggests that this is evidence of no effect. In fact, *if the studies he counted were credible (which they are not)*, his 27% figure would show the opposite. To see this, one must consider the definition of statistical significance. An effect is statistically significant when the observed effect is unlikely to have occurred by random chance. In social science (and the studies in question), the term "statistically significant" means that there is less than a 5 percent chance that the observed effect would have occurred if the true effect were zero. A study will find an effect statistically significant when the chance of observing the effect by random chance is less than five percent. By definition, if there is no effect, then approximately 5 percent of studies will be statistically significant. If studies follow standard normal distribution (which is commonly assumed), because a study could find a positive significant effect or a negative significant effect, a study will find a statistically significant positive effect when the chance of observing the effect by random chance is less than 2.5 percent.[2] As such, by definition, if there were no effect, it is unlikely that more than 2.5 percent of studies will be statistically significant and positive (see Figure 1 below). This is the relevant benchmark to compare the share of significant positive effects to, not 50%.

[2] Note that Dr. Rivkin is unclear about what probability value is used to determine statistical significance. That is, he uses the term "*statistically significant*" but does not define what level he is using to make that determination. The use the term "*statistically significant*" without specifying the criteria for determining significance (5%, 10%, one-sided tests, two-sided tests) is completely uninformative. An estimate that is statistically significant at the 10% level is not statistically significant at the 5% level, and an estimate that is statistically significant at the 5% level is not statistically significant at the 1% level. Rivkin, has failed to define the standard of evidence he used to determine significance -- this is not good scientific practice. For the purposes of this rebuttal, I assume he follows convention and uses 5 percent for a two-sided test.

Figure 1: *The probability of finding a statistically significant effect when there is no effect (using a 5% significance level)*



As discussed above, if one had a random sample of studies (and the estimated effects follow a standard normal distribution), if there were no relationship between school spending and student outcomes, we would expect to see approximately 2.5 percent of studies with a positive and statistically significant effect.[3] Any number greater than 2.5 percent would be suggestive of a real positive association, and any number considerably greater than 2.5 percent would be highly indicative of a positive effect. The 27 percent reported is more than 10 times more than 2.5%. This is considerably greater than 2.5 percent. That is, if the studies to which Dr. Rivkin refers were credible, they would provide compelling evidence that school spending improves student outcomes. To put this another way, if there were no real relationship, the probability that a single study would be statistically significant and positive is 2.5% or 0.025. If it were true that there is no relationship between school spending and student outcomes, the probability that out of 163 studies, 27 percent (44) of them would be positive and significant is less than one in 100 million. In sum, Dr. Rivkin's conclusions are not based on solid statistical reasoning and are not consistent with his own analysis.

While the discussion above highlights the flaw in Dr. Rivkin's statistical reasoning, it ignores the negative and significant studies. That is, it is worth noting that Dr. Rivkin reports that 7 percent of studies find negative and significant effects, which is also greater than 2.5 percent. As such, one *could* argue that the discussion above

[3] I say approximately because statistics is about probabilities not exact numbers.

is not proof of positive effects (*NOTE that Dr. Rivkin does not argue this*). That is, one could argue that there are more significant effects (both positive and negative) than one would expect by random chance, so that the 2.5% threshold may not be the appropriate benchmark. In statistical terms, the studies may follow a fat-tailed distribution (as opposed to a standard normal distribution). Taking this potential response seriously, one could then make a symmetry argument (that does not require specifying the size of the tails of the distribution). That is, if there is no effect of school spending on outcomes, there should be a roughly equal number of significant positive effects as significant negative effects. A test of symmetry is intuitive and straightforward and is valid even in cases in which the shape of the sampling distribution is not perfectly normal. **In fact, there are more than 3.5 times as many positive and significant effects than negative and significant.** The chance of having such a skew toward positive effects by random chance is less than one in 10,000. That is, even the most charitable application of statistical reasoning leads one to conclude that the existing studies evaluated by Dr. Rivkin (even if they were reliable, which they are not), indicate that school spending improves student outcomes.

A similar conclusion was reached in Greenwald, Hedges and Laine (1996)[4] who conduct a formal analysis of the older studies examined by Dr. Hanushek and colleagues. Using appropriate statistical tests on the same studies as explored by Dr. Hanushek and colleagues, the "analysis found that a broad range of resources were positively related to student outcomes, with effect sizes large enough to suggest that moderate increases in spending may be associated with significant increases in achievement." (abstract) Based on statistical tests (rather than assertion), they find that Dr. Hanushek and colleagues mischaracterize the old studies.

In sum, the claim that "However, a large body of research on both the effects of overall spending and specific inputs based on US data fails to find a systematic relationship between the quality of schooling as measured by the school's contribution or value-added to achievement, educational attainment or future earnings (referred to henceforth as achievement) and resources" is demonstrably false.

[4] https://www.jstor.org/stable/1170528?seq=1#metadata_info_tab_contents

The criticism of the Jacskson, et. al. and LaFortune et. al. papers does not support Dr. Rivkin's conclusion.

Referring to LaFortune et al. (2018) and Jackson et al. (2018), Dr. Rivkin states in his initial report that "the methods used in these two studies do not justify treating their findings as more compelling than the numerous other papers in this literature." (Page 4) In his revised report, he notes that his reference to Jackson et al (2018) should have been to a different paper, Jackson et al (2016). (Page 2) Either way, that statement is not consistent with the accepted standard of evidence discussed above.

In his report, Dr. Rivkin claims that the two studies are at odds with each other. They are not. Jackson et al. (2016) study the equity-based school finance reforms between 1972 and 1990, while LaFortune et al. (2018) study the adequacy-based reforms that occurred after 1990. The two studies examine different time periods, study different kinds of reforms (adequacy versus equity), and examine different outcomes (test scores verss wages). Despite differences in the context of the two studies, both studies find that school spending improves the outcomes of low-income children. Also, even if the two studies are imperfect (note that no study is perfect), that is not a reason to privilege studies that common wisdom agrees are less reliable. Dr. Rivkin has not engaged with the new literature on the topic that uses the most credible research designs available. He does not speak to numerous other studies that use credible methods, such as the 32 *other* studies that are included in the meta-analysis in my report. This ignoring of the new evidence is without any justification and warrants some justification. It appears that Dr. Rivkin only examines evidence that is consistent with his conclusions.

Dr. Rivkin claims that "Evidence in Hanushek, Rivkin and Taylor (1996) also suggests that the measurement of expenditures or resources at the state level may increase susceptibility to bias from factors not considered in the analysis," and then uses that state level conclusion to call the results in LaFortune (2018) and Jackson (2016) into question. This makes no sense because both LaFortune (2018) and Jackson (2016) study school spending at the school district level. The relevant text from the Jackson study clearly shows that we measure school spending at the school district level.[5] Dr. Rivkin is simply wrong on this point.

[5] "We compiled data on school spending, linked them to a database describing various SFRs, and linked these data to a nationally representative longitudinal data set that tracks individuals from childhood into adulthood. Education Funding data come from several sources that we combine to form a panel of per pupil spending in U.S. school districts in 1967 and annually from 1970 through 2010." See Jackson, Johnson and Persico, p. 163.

Delaware specific statements by Dr. Rivkin and Dr. Springer are wrong

Any evaluations of associations in Delaware between school spending and student achievement that are not based on quasi-experimental or other credible methodology cannot be taken as causal, and are uninformative about the effect of school spending on student outcomes in Delaware. (See discussion above about standard of evidence.) As such, none of the patterns presented by Dr. Rivkin regarding student achievement and per-pupil spending in Delaware is evidence for or against the notion that a policy to increase school spending would improve student outcomes.

Much of the analyses by Dr. Springer and Dr. Rivkin do not account for students with disabilities. As I show in my report, if one does not account for this, state spending may appear to be progressive. However, accounting for that shows that this is not the case. Also, some of the analyses by Dr. Springer and Dr. Rivkin focuses on overall school spending. The issue at hand involves spending by the State of Delaware. As such, my analysis focuses on state spending.

Dated: May 22, 2020

By: __/s/ C. Kirabo Jackson____