Comment on "New Evidence on the Heckman Curve" by David Rea and Tony Burton

Journal:	*Journal of Economic Surveys*
Manuscript ID	JOES-20-01-007
Wiley - Manuscript type:	Original Article
Keywords:	WSIPP, program evaluation, skill formation, internal rate of return

Abstract:	The paper by Rea and Burton is founded on a semantic confusion that marginal productivity is the same as an internal rate of return. This confusion plagues portions of the empirical literature in labor economics but has long been clarified in the research literature. This comment outlines why important it is to distinguish a marginal productivity from an internal rate of return, and more generally to define terms precisely. The econometric evidence to date on the productivity of life cycle skill investment supports the results from the original article. The WSIPP evidence cited by Rea and Burton is irrelevant for the purpose of testing the validity of the Heckman Curve.

The paper by Rea and Burton tabulates estimates of benefit-cost ratios by age of many social programs generated by the Washington State Institute for Public Policy (WSIPP). It uses these tabulations in an attempt to test the validity of the Heckman Curve (Heckman, 2008) that is displayed in Figure 1 of their paper.

Their paper misinterprets what the Curve is all about. Their misunderstanding, however, is common in parts of the literature, so I welcome the opportunity to exposit what my work actually shows. In addition, they draw uncritically on WSIPP estimates and appeal to a variety of "authorities" to justify their validity, but do not conduct any independent validation of those estimates.[1] They discuss the WSIPP methodology in a cursory fashion, and do not recognize many limitations of the reported benefit-cost ratios.[2] They appear not to have examined in detail any of the 300-plus studies tabulated by WSIPP.

I want to avoid an extensive discussion of the WSIPP methods and sources. The key point is that the WSIPP estimates used by Rea and Burton are irrelevant for testing the validity of the Curve. I also want to use this opportunity to warn readers about the dangers of the pseudoscience of meta-analysis, which is described as an authoritative practice by Rea and Burton. Meta-analysis replaces substantive rigorous comparisons of studies with arbitrary statistical procedures, ignoring potentially important aspects of how the reported effects are generated.

My response is in three parts. Part 1 discusses what the author of the Heckman Curve said in numerous articles and is summarized most precisely in Heckman (2008). Part 2 discusses problems that plague meta-analytic studies. Part 3 discusses the WSIPP estimates and how they are used by Rea and Burton.

1 The Heckman Curve is a Productivity Curve

What does the Heckman Curve show? It takes the perspective of a social planner examining where in the life cycle of a child to invest *in skills* to achieve the highest return, where return is defined as a *marginal product*. It plots the *marginal productivity* of the *first unit* of investment in a child at each stage of the life cycle (this is what Figure 1 used by Rea and Burton was intended to show).[3]

The logic for the Curve is simple. Skill begets skill. It creates skills that enhance later-life investment. A more skilled and motivated adolescent generally has a higher productivity of investment. For example, rates of return (marginal productivities) to college graduation are very high for the most able and motivated[4]. A major finding of the recent literature[5] is that at earlier ages the productivity of investment is higher than at later ages. Why? Building the skill base at a young age makes later investment more productive.[6]

Economists call this phenomenon *dynamic complementarity*.[7] This concept formalizes the notion that skill begets skill. The curve shows that effect. It's precisely because future investment is made more productive by early investment that early investment is so productive. This has been verified in many rigorous econometric studies that Rea and Burton ignore.

None of this says that the benefit-cost ratios or internal rates of return are necessarily higher for *all* younger-age interventions. Obviously a high productivity program with enormous cost would have a low internal rate of return, as would a low-quality program. The Curve is a *technological frontier* across programs (best practice) and not an average across all programs, however poorly executed, which is what Rea and Burton report in their study. Economists work with frontier (highest productivity) technologies in doing optimality and efficiency analyses. It turns out that ABC and Perry were high-quality skill enhancement programs and they also have high internal rates of return, but there are plenty of low-quality programs out there.

Rea and Burton misrepresent my analysis and an entire literature by equating "productivity" with the "internal rate of return." In fact, at a social optimum, the marginal productivity of a program would be equal to the marginal social cost of the program.[8]

However, many zealous advocates have interpreted my analysis as they do. Advocates should be more cautious in taking a productivity by age (at a given level of baseline skills and environment) as an internal rate of return, especially one that does not control for pre-program skills and environments.

In summary, Rea and Burton confuse two concepts: the marginal productivity of investment at one stage of the life cycle, holding constant all skills and abilities, and the rate of return (internal rate of return) to that investment accounting for costs. This semantic confusion is rife in the advocacy literature.[9]

Rea and Burton attack a strawman of their own creation. I have never equated marginal productivity with the internal rate of return, although they claim I did. They repeat what the literature has shown long ago –

that "marginal productivity" and "internal rate of return" are very different concepts. A reminder of this well-known point is useful, but scarcely deserves publication.

2 Meta Analysis of Internal Rates of Return

This technique is widely and uncritically used to summarize very diverse studies. It uses statistics to avoid the hard work of carefully comparing studies or, better yet, re-estimating the studies in a common framework.[10]

I know from personal experience how hard it is to carefully read and replicate studies and place them on a common footing. In a recent study, my coauthors and I synthesized the evidence from four leading early childhood programs evaluated by random assignment, analyzing the primary data from each in a common statistical and economic framework[11]. We verified (a) the comparability of measures used, (b) the populations targeted, (c) the outcomes measured, and (d) other relevant aspects of these programs. It was an arduous task. Neither WSIPP nor Rea and Burton conduct such rigorous analyses. WSIPP and Rea and Burton use statistical procedures in an attempt to make up for their lack of understanding of the details of each program in the WSIPP database.

Rea and Burton do not describe or justify the WSIPP analyses in any detail, but instead appeal to the authority of panels of scholars who did not carefully analyze any of the 300-plus studies WSIPP draws on. They don't do the hard work required for conducting a serious synthesis.

WSIPP is trying hard, but has a long way to go, as they admit in the technical report cited by Rea and Burton. Its efforts should be encouraged. Uncritical use of their evidence should not.

Skill enhancement programs differ greatly in terms of target populations, the quality of evaluation, evaluators and the delivery of interventions, and the scope and goal of programs. More specifically, an accurate synthesis of different programs should consider:

(a) Their target populations;

(b) The precise intervention being conducted (curricula);

(c) The length of follow up and methods used to control for cohort effects for programs with extrapolated effects using synthetic cohorts;

(d) The quality of the training of the personnel conducting the intervention;

(e) The rigor of the methodology used to evaluate each intervention;

(f) Whether control groups were contaminated (this includes the nature of the contamination and what the return are to the contaminating intervention);[12]

(g) The choice of the particular evaluation used to conduct the cost-benefit analysis of a program. Multiple evaluations are often performed for the same program. Which one was used and why are they not discussed program-by-program? WSIPP subjectively evaluated papers without clearly announced selection rules;

(h) How independent are the evaluators from the original program implementors? Do they have "skin in the game"?

(i) How are competing studies of the same program synthesized and utilized?

(j) Which methodologies are used to evaluate the program?

(k) Length of follow up;

(l) The quality, scientific status and track record of the evaluators;

(m) The baseline skills of participants;

(n) The quality of management of the program being evaluated;

(o) The quality of measures of skills used (comparability) across programs;

(p) The strength of the data base used to make "linked" estimates;[13]

(q) The methods used to monetize benefits.

Some of this heterogeneity is avoided by using benefit-cost ratios. But many items on this list remain after using benefit-cost ratios. Meta-analysis ignores many of these "details." The implicit case for meta-analysis – a synthesis of identical studies of the same treatment on the same populations with the same measurement instruments – clearly fails in the WSIPP samples.

3 The Evidence Cited

In discussing the evidence advanced by Rea and Burton, I am required to evaluate their WSIPP data source, which they use uncritically. This is an unwelcome task because I admire the ambition and intellectual honesty of the WSIPP project and recognize the daunting task it faces. At the same time, it's important to recognize the limitations of the current WSIPP study beyond those discussed in the WSIPP technical report and partially reiterated by Rea and Burton.

WSIPP attempts to address *some* of the sources of heterogeneity previously discussed. In particular, it attempts to adjust for: (i) whether independent evaluators were used (point (h) above); (ii) the quality of measures used (o); (iii) whether or not "university-based scholars" conducted or evaluated a program, an apparent liability (!); (iv) whether the control group is contaminated *by participants in the program being evaluated, but not necessarily competing programs* (part of point (f) above).[14]

Instead of controlling simultaneously for the limited set of factors they consider, it adjusts for each factor in sequential procedure with no rigorous justification for doing so.[15] A better way is to run regressions of reported effects sizes Y on adjustment factors A:

$$Y = A\beta + \varepsilon$$

where $E(\varepsilon) = 0$, and ε is orthogonal to A, and work with adjusted $Y : Y - A\hat{\beta}$. Equally without rigorous foundation (or clear documentation) is the methodology WSIPP uses to extrapolate effect sizes of the studies it summarizes outside of the sample period of the studies it utilizes. The WSIPP technical report looks authoritative at a quick glance, but on a close reading contains a lot of ad hoc procedures, which they claim are robust when subject to a "sensitivity analysis" in their Monte Carlo procedures to generate benefit-cost ratios.

As Rea and Burton acknowledge, no measure of statistical precision is reported for their estimates. We don't know which, if any, of the 300-plus WSIPP estimates are statistically significantly different from zero or statistically different from other programs. Statistical inference is being made without crucial sampling statistics. Rea and Burton contrast their analysis of WSIPP with a *single cost-benefit study* of the HighScope/Perry Preschool Program with long-term follow-up.[16] They apparently prefer estimates from an average of programs with different curricula and different target groups of varying quality to estimates from a high quality program. Their table in Appendix 2 demonstrates the high quality of Perry study in contrast with their collection of meta-analyzed studies.

They ignore a comprehensive long-run analysis of the Abecedarian (ABC) program (García et al., 2018) that includes evaluation of health benefits and carefully considers conditions required to accurately extrapolate estimates. The study generated a post-tax internal rate of return of 13.7% per annum. They also ignore the work of Elango et al. (2016), or the studies of Cunha et al. (2006), and Currie and Almond (2011).

They dismiss my work with Kautz et al. (2014) that examines the twelve studies listed in their Table 1 as being selective and compare our study with WSIPP's greater volume of studies. Our report *was* selective. It focused on high quality studies with long term follow up conducted by scholars or groups with good reputations for conducting evaluation research. This was our method of synthesizing evidence from leading-edge studies. The WSIPP studies are far less selective in many ways.

Kautz and I followed the practice in economics of examining the studies at or near the technology frontier. WSIPP also uses subjective criteria to determine which studies to include, but is far less clear in justifying its choices. The sheer number of studies they synthesize makes such nuanced analysis difficult, if not impossible. The WSIPP sausage grinder produces a crude product. Rea and Burton prefer quantity over quality.

A close inspection of the WSIPP estimates should raise eye brows. For example, their estimate of a benefit-cost ratio of 80 for a "growth mindset" intervention is grossly implausible. In fact, Kautz and I (2014) combed the literature and found only one credible evaluation of that program with long-term follow-up which we report. The benefit-cost ratio for it is far below 80. Meta-analysis is no substitute for the painstaking task of reading, reviewing, and synthesizing studies one at a time.

Not only is the internal rate of return irrelevant for estimating the Heckman Curve, the WSIPP data are far less dispositive on the internal rate of return of any of particular study. They create broad categories and squeeze programs into them and crank the meta-analysis sausage grinder to generate benefit-cost ratios. Reporting WSIPP estimates by age throws away all the nuance of program variety previously listed, including the type of services being provided and target populations. It also does not provide practitioners with a guide to best practices, or an understanding of mechanisms, which make programs work, as in Heckman et al. (2013) or Cunha et al. (2010). Understanding which mechanisms promote success is crucial for public policy analysis because it's unlikely that any particular program should be religiously copied in future evaluations. Knowledge evolves through practice, and adaptation of best practice to new situations.

Many programs are mismanaged. Many program effects depend on the quality of participants at baseline – the fundamental insight used in the Heckman Curve, since earlier investments crucially affect them. Policy makers need advice on best practice, not on average practice.

4 Summary

The paper by Rea and Burton is founded on a semantic confusion that a marginal productivity is the same as an internal rate of return. This confusion plagues portions of the empirical literature in labor economics but this has long been clarified in the research literature. Rea and Burton mischaracterize what I said and wrote and what has been supported in a large literature. Their paper, and my discussion of it, serves as a reminder of how important it is to distinguish a marginal productivity from an internal rate of return, and more generally to define terms precisely and to read papers carefully.

The econometric evidence to date on the productivity of life cycle skill investment strongly supports the Curve.[17] The WSIPP evidence – unstandardized for many relevant determinants of effect sizes – is irrelevant for the purpose of testing the validity of the Curve. The Heckman Curve is alive and well. Of course, future studies may disprove it. That is the nature of science.

I strongly applaud efforts like those undertaken by WSIPP, but at the same time, warn readers about the dangers of meta-analysis. There are numerous prior critiques of meta-analysis, see e.g. Anderson and Kichkha (2017); Feinstein (1995); Eysenck (1978). It is a lazy person's way to evaluate and compare programs. Uncritical use of statistical meta-analysis will cause more harm than good in evaluating and guiding policy.

Notes

[1]The WSIPP technical report was prepared by a government agency and was not published in a refereed journal.

[2]They refer the reader to a 220-page technical report issued by WSIPP (2018). It lacks important details and rigorous justification of the choice of studies estimated, the range of parameter values assumed, and the methodology for extrapolating estimates out of sample. Portions of this methodology were updated by WSIPP after the estimates used by Rea and Burton were generated.

[3]See Heckman (2008), pp. 309-314, and especially Figures 18a and 18b where this is made crystal clear.

[4]See Cameron and Heckman, 2001. In Heckman (2008), I explicitly write "Evidence is presented in this paper that high quality early childhood interventions foster abilities and that inequality can be attacked at its source". In Section VII, I summarize the experimental evidence that test scores and adult achievement can be improved by high quality intervention.

[5]Cunha et al. (2010); Del Boca et al. (2019); Agostinelli and Wiswall (2016); Cunha and Heckman (2008); Del Boca et al. (2014)

[6]See Heckman and Mosso, 2014 for further discussion and more evidence.

[7]See Cunha et al. (2010) and Cunha and Heckman (2007).

[8]The Heckman Curve is an out-of-equilibrium concept.

[9] It is also common in the applied labor economics literature. The "Mincer return" to schooling is, in fact, the marginal product of schooling. Under special conditions delineated in Heckman et al. (2006), it is a true internal rate of return. However, as shown in that paper, those conditions do not apply in U.S. data. Going more deeply, the internal rate of return is itself a flawed measure of true economic rates of return as it fails to capture continuation and option values. See Eisenhauer et al. (2015) and Heckman et al. (2006).

[10]Rea and Burton stress that WSIPP use a common framework, however, WSIPP does not analyze the primary data of each study used with a common analytic framework.

[11]Elango et al. (2016). Rea and Burton ignore this study.

[12]See the evidence on the importance of doing so in Heckman et al. (2000) and Kline and Walters (2016).

[13]See García et al. (2019).

[14]See Heckman et al. (2000) and Kline and Walters (2016) for the discussion of the importance of contamination bias in evaluations.

[15]Stephanie Lee informs me WSIPP has improved the adjustment analysis beyond that used to generate the estimates of WSIPP utilized by Rea and Burton. It is still sequential rather than simultaneous, however.

[16]Heckman et al., 2010 who show that program has a 7-10% annual rate of return.

[17]See, e.g., Cunha et al. (2010). That paper won the 2013 Frisch Prize as the best applied paper published in *Econometrica* over the period 2010-2013. Numerous other papers build on this that are cited in Section 1.

References

Agostinelli, F. and M. Wiswall (2016, July). Identification of dynamic latent factor models: The implications of re-normalization in a model of child development. Working Paper, W. P. Carey School of Business, Arizona State University.

Anderson, R. G. and A. Kichkha (2017, May). Replication, meta-analysis, and research synthesis in economics. *American Economic Review 107*(5), 56–59.

Cameron, S. V. and J. J. Heckman (2001, June). The dynamics of educational attainment for black, Hispanic, and white males. *Journal of Political Economy 109*(3), 455–499.

Cunha, F. and J. J. Heckman (2007, May). The technology of skill formation. *American Economic Review 97*(2), 31–47.

Cunha, F. and J. J. Heckman (2008, Fall). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources 43*(4), 738–782.

Cunha, F., J. J. Heckman, and S. Navarro (2006). Counterfactual analysis of inequality and social mobility. In S. L. Morgan, D. B. Grusky, and G. S. Fields (Eds.), *Mobility and Inequality: Frontiers of Research in Sociology and Economics*, Chapter 4, pp. 290–348. Stanford, CA: Stanford University Press.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010, May). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica 78*(3), 883–931.

Currie, J. and D. Almond (2011). Human capital development before age five. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 4B, Chapter 15, pp. 1315–1486. Amsterdam: Elsevier Science B. V.

Del Boca, D., C. J. Flinn, E. Verriest, and M. J. Wiswall (2019). Actors in the child development process. Technical Report 25596, NBER Working Paper.

Del Boca, D., C. J. Flinn, and M. Wiswall (2014). Household choices and child development. *Review of Economic Studies 81*(1), 137–185.

Eisenhauer, P., J. J. Heckman, and S. Mosso (2015, May). Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments. *International Economic Review 56*(2: Special Issue: Estimation of Dynamic Stochastic Models in Empirical Microeconomics: In Honor of Kenneth I. Wolpin), 331–357.

Elango, S., , J. L. García, J. J. Heckman, and A. Hojman (2016). Early childhood education. In R. A. Moffitt (Ed.), *Economics of Means-Tested Transfer Programs in the United States*, Volume 2, Chapter 4, pp. 235–297. Chicago: University of Chicago Press.

Eysenck, H. J. (1978, May). An exercise in mega-silliness. *American Psychologist 33*(5).

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology 48*(1), 71 – 79. The Potsdam International Consultation on Meta-Analysis.

García, J. L., J. J. Heckman, D. E. Leaf, and M. J. Prados (2018). Quantifying the life-cycle benefits of a prototypical early childhood program. Forthcoming at the *Journal of Political Economy*, 2020.

García, J. L., J. J. Heckman, and A. L. Ziff (2019). Early childhood education and crime. *Infant Mental Health Journal 40*(1).

Heckman, J. J. (2008, July). Schools, skills, and synapses. *Economic Inquiry 46*(3), 289–324.

Heckman, J. J., N. Hohmann, J. Smith, and M. Khoo (2000, May). Substitution and dropout bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics 115*(2), 651–694.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Volume 1, Chapter 7, pp. 307–458. Amsterdam: Elsevier.

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Q. Yavitz (2010, February). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics 94*(1–2), 114–128.

Heckman, J. J. and S. Mosso (2014, August). The economics of human development and social mobility. *Annual Review of Economics 6*(1), 689–733.

Heckman, J. J., R. Pinto, and P. A. Savelyev (2013, October). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review 103*(6), 2052–2086.

Kautz, T., J. J. Heckman, R. Diris, B. ter Weel, and L. Borghans (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. Technical report, Organisation for Economic Co-operation and Development, Paris. Available at https://www.oecd.org/edu/ceri/Fostering-and-Measuring-Skills-Improving-Cognitive-and-Non-Cognitive-Skills-to-Promote-Lifetime-Success.pdf.

Kline, P. and C. Walters (2016). Evaluating public programs with close substitutes: The case of Head Start. Quarterly Journal of Economics *131*(4), 1795–1848.

Washington State Institute for Public Policy (2018). Benefit-cost technical documentation. Technical report, Olympia, WA.