HCEO WORKING PAPER SERIES

"Unsettled Science on Longer-run Effects of Early Education:" A Response

Jorge Luis García
James J. Heckman

Working Paper

2024-012

06/2024

HUMAN CAPITAL AND
ECONOMIC OPPORTUNITY
GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

"Unsettled Science on Longer-run Effects of Early Education:" A Response

Jorge Luis García,[1] James J. Heckman[2*]

[1]Department of Economics, Texas A&M University
2935 Research Parkway, College Station, TX 77845, USA
jlgarcia@tamu.edu

[2]Department of Economics, University of Chicago
1126 East 59th Street, Chicago, IL 60637, USA
jjh@uchicago.edu

Abstract

A recent comment published in Science argues that the evidence on the long-term effects of early childhood education is unsettled. We qualify this comment and contrast it with comprehensive studies based on established principles of scientific practice. Burchinal et al. (2024) base their assessment on flawed experimental evaluations. They mischaracterize the state of knowledge by selectively evaluating evidence and ignoring rigorous, long-term studies based on the Perry Preschool and Carolina Abecedarian Projects. High-quality early childhood education programs achieve consistent long-term benefits when proper controls and standardizations are applied. An essential mechanism for their success is fostering parental investment and effective parenting. We underscore the necessity of mechanism-focused research to guide early childhood education policies. Well-conducted studies demonstrate the long-term effectiveness of high-quality early education programs, including Head Start.

1. Introduction

Burchinal et al. (2024) discuss the "unsettled science" on the longer-run effects of early life interventions aiming to foster human capital. Their comment casts doubt on a substantial body of evidence they partially and incompletely assess; it contrasts with the comprehensive assessments in Elango et al. (2016) and García and Heckman (2023) based on analyses of primary data and, more generally, on established principles of essential scientific practice (e.g., replication, inference, a quest for commonalities, coherence across studies, and understanding of the mechanisms producing impacts).

We agree with Burchinal et al. (2024) that knowing more about any phenomenon is always desirable, whether it is early life human capital or the physics of dark matter. We disagree with the authors in their characterization of the state of knowledge in the literature. Our main disagreement is not about their analysis of any particular program, although we have reservations with their interpretation of what constitutes the science of child development. The authors aim to find "the best" program among a basket of programs with "good" long-term effects. They never define the long run. They lump diverse programs into the common category of "early interventions." They fail to acknowledge the fifty-year follow-up of the Perry Preschool Project (Weikart et al., 1978) with substantial, statistically significant beneficial effects for the original participants and their children that survive rigorous small sample analyses. They fail to acknowledge another rigorous body of long-run evidence based on the Carolina Abecedarian (ABC) Project (Ramey and Smith, 1977).[1]

Science is about understanding mechanisms. García and Heckman (2023) analyze the common mechanisms operating in several programs evaluated by random assignment around the world, including Perry and ABC. That analysis revisits the long-run (through age 35) impact of a low-cost Jamaican home visiting program (Grantham-McGregor et al., 1991)

[1]Instead of providing lists of references, we refer readers to Elango et al. (2016) and García and Heckman (2023) for a discussion of the numerous studies of Perry and ABC. Studies include analyses of long-run treatment effects, cost-benefit analyses, and analyses of treatment-effect mechanisms.

operating in the slums of Kingston, and a variety of versions of that approach launched in diverse settings with different contexts and social supports.[2]

The recurring finding in García and Heckman (2023) is that programs that promote active parenting—particularly, attachment of children with parents and warm, supportive parenting styles—are effective in diverse contexts. A vast body of scientific literature supports this conclusion. Effect sizes on cognitive and non-cognitive skills are comparable across programs. This is a solid finding that should guide the design of future programs; it should come as no surprise given the universality across peoples and time of the process of child development documented in Ertem et al. (2018) and Fernald et al. (2017).

2. Comments on the State of Evidence in the Literature

Burchinal et al. (2024) is based, in part, on evidence from a flawed experimental implementation of Head Start, the Head Start Impact Study (HSIS), and a state program implemented in Tennessee, within an experimental implementation known to have a flawed design. Burchinal et al. (2024) discards evidence from Perry and ABC and fails to consider the Infant Health and Development Program. This program is evaluated by a large-scale, multisite, randomized controlled trial that shows effectiveness at improving outcomes up to age eighteen (Dougan et al., 2023).

Burchinal et al. (2024) mischaracterizes findings and omits basic features of other programs they use as evidence. Their discussion of HSIS is an example. HSIS is known to be flawed by substitution bias. The parents of control-group children used alternatives comparable to Head Start; in many cases, these alternatives were other Head Start centers. Correcting for this bias, Kline and Walters (2016) show that treatment-group children substantially outperform control-group children in achievement tests taken a year or two after the program implementation, in contrast with the initial reports (Puma et al., 2010), on

[2]Examples include programs in Bangladesh, Brazil, Colombia, China, India, and Ireland discussed in García and Heckman (2023).

which Burchinal et al. (2024) relies.

An adequate characterization would be that Head Start is a large-scale early childhood program that implements the ingredients of Perry and ABC in many of its programs. The analysis of HSIS by Kline and Walters (2016) actually establishes that Head Start is effective and has a high rate of return. Multiple other studies have shown that Head Start and similar programs effectively improve socioeconomic outcomes at different life-cycle stages. For extensive discussion, see Elango et al. (2016).

Another analysis of a program that is fundamental to the argument of Burchinal et al. (2024) is the Tennessee Voluntary Prekindergarten Program (TN-VPK), a statewide kindergarten program targeting disadvantaged four-year-old children one year before kindergarten. TN-VPK began as a pilot program in 1998 and became statewide in 2005. More details on its implementation, quality, and funding are in Elango et al. (2016). TN-VPK shares many features in common with Head Start, and the authors use evidence from it as evidence against the effectiveness of programs like Head Start.

A randomized control trial evaluated TN-VPK. However, the evaluation has major flaws, and, as in the case of HSIS, the interpretation of its results is clouded by the presence of control contamination and selective attrition. Program implementers requested parental consent *after* performing the randomization, causing substantial selective attrition from the study. For the first cohort of participants, only 46% of the parents in the treatment group consented to enter the study, and 32% of the parents in the control group consented. The consent rates for the second cohort were 74% for the treatment group and 68% for the control group. This sampling created a major problem of selective attrition. Straightforward experimental methods to evaluate this program are invalid. The evaluators thus rely on non-experimental methods (Lipsey et al., 2015, 2013). The evaluation of TN-VPK used a representative sample of all the programs in the state. In their sample, 27% of the children in the control group attended Head Start or a private, center-based preschool program (Lipsey

et al., 2015). Furthermore, some centers in the program received Head Start funding, effectively making them Head Start centers. Therefore, current evaluations of TN-VPK do not address confounds or identify clear counterfactuals.

The evaluation of TN-VPK does not represent strong evidence against the effectiveness of early childhood education programs. Instead, it illustrates that interpreting effects without accounting for flaws in the experimental design or estimating clear counterfactuals produces misleading policy conclusions. HSIS and TN-VPK reveal the folly of blind faith in the validity of randomized control studies.

Burchinal et al. (2024) ignore how essential Perry and ABC have been to the scientific understanding of the life-cycle impacts of early education. These programs have been analyzed over the long run (age 54 for Perry and age 45 for ABC). These programs' strong positive long-term results are supported by a wide variety of robustness studies discussed in García and Heckman (2023). Some analysts object to the small sample sizes of these programs. In response to such objections, studies have developed and applied tools for small sample statistical inference that have been refereed in rigorous econometrics journals and support positive conclusions across a variety of methodologies (Heckman and Karapakula, 2021; Heckman et al., 2010, 2024).

Because small samples can be affected by outliers, studies of Perry and ABC conduct extensive robustness analyses, focusing on various subsamples of data. Small sample exact inference procedures are used, and corrections are made for multiple hypothesis testing. Claims that small samples produce inflated treatment-effect estimates are without foundation. Instead of reporting a disjointed collection of treatment effects, García et al. (2020) and Bennhoff et al. (2024) aggregate the effectiveness of each program using a single economic and policy-relevant statistic—the benefit-cost ratio of each intervention and its components.

García and Heckman (2023) examines the representativeness of Perry and ABC; it notes that the samples of Perry and ABC are representative of the disadvantaged, low-

income African American populations that would benefit from early intervention. Precisely, they indicate that the percentage of children eligible for Perry today is 10%. This is the percentage of children born in households satisfying the eligibility criteria to participate in the program in the most recent US census. García and Heckman (2023) argue that the similarity of eligibility criteria across programs implies that a similar percent of children in the US is eligible for ABC. Therefore, the findings from Perry and ABC have immediate policy relevance and applicability.

3. Summary

Burchinal et al. (2024) confuse statistics with science. They break a fundamental rule of science and basic logic—comparing the comparable—and instead compare an assortment of programs based on different combinations of unspecified developmental mechanisms, using a variety of often incomparable measures on diverse populations that usually differ significantly in the home lives supporting children. Any evidence analyzed in this fashion is "unsettled" because of the various uncontrolled features in the comparison. Relying on flawed studies of programs to question the validity of an entire literature is circular logic. When proper controls and standardization are made, the evidence from Perry, ABC, and Head Start indicates that high-quality early childhood education has universal ingredients that lead to long-term effectiveness when applied in various contexts.

We fully agree that there is more to be known. However, the truly valuable knowledge is about the operation of mechanisms in standardized environments, not comparisons of collections of unspecified treatment effects based on different measurement systems, even if they are experimentally generated.

References

Bennhoff, F. H., J. L. García, and D. E. Leaf (2024). The Dynastic Benefits of Early-Childhood Education: Participant Benefits and Family Spillovers. *Journal of Human Capital 18*(1), 44–73.

Burchinal, M., A. Whitaker, J. Jenkins, D. Bailey, T. Watts, G. Duncan, and E. Hart (2024). Unsettled Science on Longer-run Effects of Early Education. *Science 384*(6695), 506–508.

Dougan, W. R., J. L. García, and I. Polovnikov (2023). High-Quality Early-Childhood Education at Scale: Evidence from a Multisite Randomized Trial. Working Paper w31694, National Bureau of Economic Research, Cambridge, MA.

Elango, S., J. L. García, J. J. Heckman, and A. Hojman (2016). Early Childhood Education. In *Economics of Means-Tested Transfer Programs in the United States*, Volume II. University of Chicago Press.

Ertem, I. O., V. Krishnamurthy, M. C. Mulaudzi, Y. Sguassero, H. Balta, O. Gulumser, B. Bilik, R. Srinivasan, B. Johnson, G. Gan, L. Calvocoressi, V. Shabanova, and B. W. C. Forsyth (2018). Similarities and differences in child Development from Birth to Age 3 Years by Sex and Across Four Countries: A Cross-sectional, Observational Study. *The Lancet Global Health 6*(3), e279–e291.

Fernald, L., E. L. Prado, P. Kariger, and A. Raiker (2017). A Toolkit for Measuring Early Childhood Development in Low and Middle-income Countries. Technical report, Washington DC: World Bank.

García, J. L. and J. J. Heckman (2023). Parenting Promotes Social Mobility Within and Across Generations. *Annual Review of Economics 15*, 349–388.

García, J. L., J. J. Heckman, D. E. Leaf, and M. J. Prados (2020). Quantifying the Life-Cycle Benefits of an Influential Early-Childhood Program. *Journal of Political Economy 128*(7), 2502–2541.

Grantham-McGregor, S., C. Powell, S. Walker, and J. Himes (1991). Nutritional Supplementation, Psychosocial Stimulation, and Mental Development of Stunted Children: The Jamaican Study. *The Lancet 338*(8758), 1–5.

Heckman, J. J. and G. Karapakula (2021). Using a Satisficing Model of Experimenter Decision-Making to Guide Finite-sample Inference for Compromised Experiments. *Econometrics Journal 24*(2), C1–C39.

Heckman, J. J., R. Pinto, S. H. Moon, A. Yavitz, and P. Savelyev (2010). Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program. *Quantitative Economics 1*(1), 1–46.

Heckman, J. J., R. Pinto, and A. M. Shaikh (2024). Dealing with Imperfect Randomization: Inference for the Highscope Perry Preschool Program. *Journal of Econometrics Forthcoming*.

Kline, P. and C. R. Walters (2016). Evaluating Public Programs with Close Substitutes: The Case of HeadStart. *Quarterly Journal of Economics 131*(4), 1795–1848.

Lipsey, M., D. Farran, and K. Hofer (2015). A Randomized Control Trial of the Effects of a Statewide Voluntary Prekindergarten Program on Children's Skills and Behaviors through Third Grade. Technical report, Nashville, TN: Vanderbilt University, Peabody Research Institute.

Lipsey, M., K. Hofer, N. Dong, D. Farran, and C. Bilbery (2013). Evaluation of the Tennessee Voluntary Prekindergarten Program: Kindergarten and First Grade Follow-Up Results from the Randomized Control Design. Technical report, Nashville, TN: Vanderbilt University, Peabody Research Institute.

Puma, M., S. Bell, R. Cook, and C. Heid (2010). Head Start Impact Study Final Report. Technical report, US Department of Health and Human Services, Washington DC.

Ramey, C. T. and B. Smith (1977). Assessing the Intellectural Consequences of Early Intervention with High-Risk Infants. *American Journal of Mental Deficiency 81*(4), 318–324.

Weikart, D. P., J. T. Bond, and J. T. McNeil (1978). *The Ypsilanti Perry Preschool Project: Preschool Years and Longitudinal Results through Fourth Grade*. Ypsilanti, MI: HighScope Press.