NIEER
NATIONAL INSTITUTE FOR EARLY EDUCATION RESEARCH

Response to The Early Learning Collaborative Act of 2013: Evaluation of the Operations and Effectiveness of the Program

Allison Friedman-Krauss, Ph.D.
National Institute for Early Education Research
Rutgers, The State University of New Jersey

PEER's report evaluating Mississippi's Early Learning Collaborative Preschool Program relies on research that does not meet the high standards needed to reach valid conclusions regarding program effectiveness that is typically used to inform policy. Several limitations of the data and methodology are acknowledged in the report. Indeed the authors of the report claim on page 37 that "a more detailed study, including methods and data collection planned specifically for the purpose would be necessary to evaluate prekindergarten programs in Mississippi in the most thorough manner." Despite this acknowledgement, the PEER report reaches the conclusion that there is a smaller increase in the percent of children meeting or exceeding kindergarten readiness standards in the Collaborative Preschool Program than in other publicly funded preschool programs. Below I highlight 7 key limitations of the methods used in this evaluation that raise doubts about the validity of the conclusions drawn, specifically in regards to the effective of the Collaborative Preschool program.

1. The PEER report evaluated the first full year of Mississippi's Early Learning Collaborative Preschool program. High quality preschool does not happen over night. Programs need time to develop. To see the "true" effects of a program, evaluations should not be completed during the first year of implementation. New Jersey's Abbott Preschool Program, which is considered a high quality early childhood education (ECE) program and has positive long-term impacts on children, was not evaluated with respect to outcomes until five years after the program began. Quality improved over the entire first decade of the program and using quality in the first year of the program as an indicator of long term quality would have been misleading and counterproductive (see Frede, Jung, Barnett, & Figueras, 2009).

2. Causal claims about the effectiveness of educational programs are difficult to make with confidence and require strong research designs (Shadish, Cook, & Campbell, 2002). In addition to randomized control trials, rigorous quasi-experimental methods like regression discontinuity designs are commonly used to evaluate ECE programs (see for example Frede, Jung, Barnett, Lamy, & Figueras, 2007; Gormley, Phillips, & Gayer, 2009; Hustedt, Jung, Barnett, & Williams, 2015; Weiland & Yoshikawa, 2013). Yet the PEER report did not use any of these rigorous methods to evaluate the Early Learning Collaborative Preschool program. Instead, it relied on descriptive or correlational analyses to compare the program to other publicly-funded preschool programs. These analyses are not well-described in the report and it seems unlikely that they provide an adequate basis for concluding that one program caused better child outcomes than the other program, and for making policy recommendations. In fact, the PEER report references the What Works Clearinghouse regarding research needed to provide evidence for the effectiveness of curricula but does not apply these same high, rigorous standards to evaluating their own methods. The report also acknowledges on page 25 that "data from the kindergarten readiness assessment do not include sufficient information with which to isolate confounding variables," a key

threat to internal validity especially when experimental or quasi-experimental methods are not used.

4. There are also problems with the comparison group used. Related to #3 above, no information is provided about baseline equivalence of the two groups. That is, there may be problems with selection bias if the baseline characteristics of children, families, and even neighborhoods differ between children who attended the two different groups. In this case, it is not possible to determine if the differences in the spring are due to the effect of the program or the difference in the two samples of children. Although this threat is mitigated somewhat by "controlling" for baseline test scores, it is not eliminated. Related, are the two sets of programs that are being compared in the same cities or neighborhoods? If not, there may be other factors driving differences in children's test scores than the receipt of Early Learning Collaborative funding. Finally, do the two programs offer similar "dosages" of ECE? Research suggests that spending more time in a center-based program is associated with larger child gains (McCartney et al., 2010; NICHD, 2003). Therefore it is important that the children attending the Collaborative preschools and those included in the comparison group attend the ECE programs for a similar number of hours per day and days per week. Yet this information is another unknown.

5. It is also concerning that only children who completed both pre- and post-test were included in the final analysis. Although this is one method for handling missing data, it relies on the strict assumption that data is missing completely at random - that each child has the same probability of having missing data (Enders et al., 2013; van Buuren, 2012). There is no evidence provided that this assumption is met. If children who did and did not have fall and spring test scores in the Collaboratives and the comparison group differed, excluding them from the analyses could bias the results. Without knowing how these children differed, it is impossible to know how excluding these children might have biased the results.

6. The PEER report compared the percent of children in the Early Learning Collaboratives and comparison preschools who changed from being below a score of 498 to above a score of 498 between the fall and spring. Although using a change score is commendable, this approach masks a lot of variation in children's test scores and changes in children's test scores. For example, if a child in the comparison group scored 496 in the fall and 500 in the spring, this child would be coded as improving. Yet if a child in the Collaborative group scored a 400 in the fall and 490 in the spring, this child would appear to not have changed at all, despite a gain of 90 points. Using the scaled score would show the size of the actual gains, which might show a different patter of results. If the threshold approach is to be used, is it fair to use the same threshold in the fall and spring of pre-K? That is, if children who are ready for kindergarten should score at least 498 in the spring of pre-K and 530 in the fall of K, shouldn't the threshold during the fall of pre-K be lower?

7. Finally, the conclusions in this report rely only on one outcome - literacy. PEER acknowledges in the report that the program should be evaluated in reference to broader domains covered in the state's Early Learning Standards for Classrooms Serving Four-Year-Old Children. Therefore, conclusions about the programs effectiveness in supporting children's learning and development in other domains cannot be drawn.