WWC Intervention Report

A summary of findings from a systematic review of the evidence

ies INSTITUTE OF EDUCATION SCIENCES

Early Childhood Education **July 2015 Revised***

Head Start

Program Description[1]

Head Start is a national, federally funded program that provides services to promote school readiness for children from birth to age 5 from predominantly low-income families.[2] These services are provided to both children and their families and include education, health and nutrition, family engagement, and other social services.

Head Start program administrators are given the flexibility to design service delivery to be responsive to cultural, linguistic, and other contextual needs of local communities, leading to considerable variability in the services offered. *Head Start* service models also vary according to family needs, such that children and families may be served through center-based or family child care, home visits, or a combination of programs that operate full or half days for 8–12 months per year.[3] This review focuses on the effects of *Head Start* programs designed for children ages 3–5. The *Head Start* programs include a variety of *Head Start* service models.

Research[4]

The What Works Clearinghouse (WWC) identified one study of *Head Start* that both falls within the scope of the Early Childhood Education topic area and meets WWC group design standards.[5] The study meets WWC group design standards without reservations, and no studies meet WWC group design standards with reservations. The study included 3,697 three- and four-year-old children in a nationally-representative sample.

The WWC considers the extent of evidence for *Head Start* on the school readiness outcomes of 3- and 4-year-old children to be small for three outcome domains—general reading achievement, mathematics achievement, and social-emotional development. There were no studies that meet standards in the five other domains, so this intervention report does not report on the effectiveness of *Head Start* for those domains.[6] (See the Effectiveness Summary on p. 5 for more details of effectiveness by domain.)

Effectiveness

Head Start was found to have potentially positive effects on general reading achievement and no discernible effects on mathematics achievement and social-emotional development for 3- and 4-year-old children.

Report Contents

This intervention report presents findings from a systematic review of *Head Start* conducted using the WWC Procedures and Standards Handbook, version 3.0, and the Early Childhood Education review protocol, version 3.0.

Table 1. Summary of findings[7]

Outcome domain	Rating of effectiveness	Improvement index *(percentile points)*		Number of studies	Number of students	Extent of evidence
		Average	Range			
General reading achievement	Potentially positive effects	+13	+12 to +14	1	3,697	Small
Mathematics achievement	No discernible effects	+3	na	1	1,617	Smal
Social-emotional development	No discernible effects	+1	–1 to +5	1	3,693	Small

na = not applicable

Program Information

Background

Head Start was first launched as an 8-week demonstration project in the summers of 1965 and 1966. Since then, *Head Start* has served more than 30 million children and grown to include a multitude of program options. *Head Start* was most recently reauthorized in 2007. Currently, the program is administered by the Administration for Children and Families in the U.S. Department of Health and Human Services. Address: Office of Head Start, 1250 Maryland Ave., SW, Washington, DC 20024. Email: HeadStart@eclkc.info. Web: http://www.acf.hhs.gov/programs/ohs. Telephone: (866) 763-6481.

Program details

Head Start is a national, federally funded program for preschool children from low-income families. *Head Start's* main purpose is to prepare children for school. The program seeks to promote school readiness by bolstering a child's development and learning. Services focus on language and literacy skills, cognition and general knowledge, physical development and health, social and emotional development, and approaches to learning. Programs may be based in centers or schools, family child care homes, and children's own homes (home visits).

Head Start services are designed to be responsive to each child and family's ethnic, cultural, and linguistic heritage. Children who attend *Head Start* participate in a variety of educational activities. They also receive free medical, dental, and mental health screenings. They are provided with healthy meals and snacks, as well as opportunities to play indoors and outdoors in a safe setting. In addition, *Head Start* programs work with children's families to assist with accessing regular health care and community resources for low-income families, and helping them actively engage in their children's development and early learning.

Head Start programs work directly with local agencies and can differ based on community needs. The Office of Head Start (http://www.acf.hhs.gov/programs/ohs) is responsible for oversight of grantees, quality assurance, and technical assistance for program staff in their delivery of services to children and families.

Cost

The cost of implementing *Head Start* is available from the Office of Head Start. As a rule, local grantees must provide a 20% cash or in-kind match to federal funds. No more than 15% of total program costs may be used for program administration. In addition, some localities offer funding to expand *Head Start* to additional children within their jurisdiction.

Research Summary

The WWC identified 40 eligible studies that investigated the effects of *Head Start* on the school readiness of preschool-aged children. An additional 50 studies were identified but do not meet WWC eligibility criteria for review in this topic area. Citations for all 90 studies are in the References section, which begins on p. 7.

Table 2. Scope of reviewed research

Grades	PK
Delivery method	Individual, Small group, Whole class, Whole school
Program type	School level

The WWC reviewed 40 eligible studies against group design standards. One study (U.S. Department of Health and Human Services, Administration for Children and Families [DHHS ACF], 2010) is a randomized controlled trial that meets WWC group design standards without reservations. The study is summarized in this report. Thirty-nine studies do not meet WWC design standards.

Summary of study meeting WWC group design standards without reservations

DHHS ACF (2010) reports on the impact of *Head Start* on the development of preschool children's school readiness skills. The authors of this study created a nationally-representative sample by first randomly identifying 84 grantee or delegate agencies[8] and 378 *Head Start* centers in 23 states to conduct the experiment.[9] Children in the comparison group experienced diverse types of early care and education settings, ranging from parent-only care to programs that were similar in type and services to more typical *Head Start* programs.

The study design and presentation of impacts focused on two cohorts of children.

Three-year-old cohort: The DHHS ACF (2010) study included 3-year-old children whose families were applying to the selected *Head Start* programs for the first time. These children were then randomly assigned either to be offered *Head Start* (1,278 children) or to be in the comparison group (784 children).[10] The authors reported that 17.3% of those assigned to the comparison group actually enrolled in a *Head Start* program that was not selected for inclusion in the study.[11]

Four-year-old cohort: The DHHS ACF (2010) study included 4-year-old children whose families were applying to the selected *Head Start* programs for the first time. These children were then randomly assigned either to be offered *Head Start* (1,008 children) or to be in the comparison group (627 children).[12] The authors reported that 13.9% of those assigned to the comparison group actually enrolled in a *Head Start* program that was not selected for inclusion in the study.[13]

The study findings and summary of effectiveness presented in the main body of this report are based on total score outcomes that were measured at the end of the children's first year in *Head Start* (in 2003) and represent the immediate effects of *Head Start*.[14] Following the prioritization of immediate outcomes as indicated in the Early Childhood Education review protocol (version 3.0), findings associated with the 4-year-old cohort outcome measures collected in spring 2004, 2005, and 2007 in the general reading achievement, mathematics achievement, social-emotional development, alphabetics, cognition, comprehension, and language development domains are presented as supplemental outcomes in Appendix D, as they represent intermediate to longer-term follow-up effects of *Head Start* at kindergarten, first grade, and third grade, respectively. Appendix D also presents subscale findings from the social-emotional development domain[15] and comparisons based on mothers' race/ethnicity in the mathematics achievement, social-emotional development, alphabetics, cognition, comprehension, and language development domains for both the 3-year-old and 4-year-old cohorts. These additional comparisons are presented as supplemental findings in the appendix and do not factor into the intervention's ratings of effectiveness.[16]

Summary of studies meeting WWC group design standards with reservations

No studies of *Head Start* met WWC group design standards with reservations.

Effectiveness Summary

The WWC review of *Head Start* for the Early Childhood Education topic area includes student outcomes in eight domains: alphabetics, cognition, comprehension, fluency, general reading achievement, language development, mathematics achievement, and social-emotional development. The one study of *Head Start* that meets WWC group design standards without reservations reported findings in three of the eight domains: (a) general reading achievement, (b) mathematics achievement, and (c) social-emotional development. The findings below present both the authors' estimates and WWC-calculated estimates of the size and statistical significance of the effects of *Head Start* on 3- and 4-year-old children. Additional comparisons are presented as supplemental findings in Appendix D and do not factor into the intervention's ratings of effectiveness.[17] For a more detailed description of the rating of effectiveness and extent of evidence criteria, see the WWC Rating Criteria on p. 50.

Summary of effectiveness for the general reading achievement domain

One study that meets WWC group design standards without reservations reported findings in the general reading achievement domain.

The DHHS ACF (2010) study investigated one outcome in the general reading achievement domain that meets WWC group design standards without reservations: The Parent Emergent Literacy Scale (PELS). The PELS measures children's literacy skills in five areas using parent ratings, which include letter recognition, counting, name writing, and primary color identification. The authors reported, and the WWC confirmed, a statistically significant and positive effect of *Head Start* on children's PELS ratings for both the 3-year-old and 4-year-old cohorts at the end of the intervention year. The WWC characterizes these study findings as a statistically significant positive effect.

Thus, for the general reading achievement domain, one study showed statistically significant positive effects, and no studies showed an indeterminate effect or a statistically significant or substantively important negative effect. This results in a rating of potentially positive effects, with a small extent of evidence.

Table 3.1 Rating of effectiveness and extent of evidence for the general reading achievement domain

Rating of effectiveness	Criteria met
Potentially positive effects *Evidence of a positive effect with no overriding contrary evidence.*	In the one study that reported findings, the estimated impact of the intervention on outcomes in the *general reading achievement* domain was positive and statistically significant.
Extent of evidence	**Criteria met**
Small	One study that included 3,697 3- and 4-year-old children in a nationally-representative sample reported evidence of effectiveness in the *general reading achievement* domain.

Summary of effectiveness for the mathematics achievement domain

One study that meets WWC group design standards without reservations reported findings in the mathematics achievement domain.

The DHHS ACF (2010) study reported one outcome in the mathematics achievement domain that meets WWC group design standards without reservations: the Counting Bears Test. The Counting Bears Test measures counting ability and understanding of one-to-one correspondence. Impacts for this outcome were only presented for the 4-year-old cohort. The authors reported, and the WWC confirmed, no statistically significant or substantively important difference between the *Head Start* and comparison groups on this measure. The WWC characterizes this study as having an indeterminate effect because the mean effect is neither statistically significant nor substantively important.

Thus, for the mathematics achievement domain, one study showed an indeterminate effect, and no studies showed a statistically significant or substantively important positive or negative effect. This results in a rating of no discernible effects, with a small extent of evidence.

Table 3.2 Rating of effectiveness and extent of evidence for the mathematics achievement domain

Rating of effectiveness	Criteria met
No discernible effects *No affirmative evidence of effects.*	In the one study that reported findings, the estimated impact of the intervention on outcomes in the *mathematics achievement* domain was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Small	One study that included 1,617 4-year-old children in a nationally-representative sample reported evidence of effectiveness in the *mathematics achievement* domain.

Summary of effectiveness for the social-emotional development domain

One study that meets WWC group design standards without reservations reported findings in the social-emotional development domain.

The DHHS ACF (2010) study reported three outcomes in the social-emotional development domain that meet WWC group design standards without reservations: the Total Problem Behavior Scale,[18] the Social Competencies Checklist, and the Social Skills and Positive Approaches to Learning Scale. The authors reported, and the WWC confirmed, no statistically significant or substantively important effects on any of these measures for either the 3-year-old or the 4-year-old cohorts. The WWC characterizes this study as having an indeterminate effect because the mean effect is neither statistically significant nor substantively important.[19]

Thus, for the social-emotional development domain, one study showed an indeterminate effect, and no studies showed a statistically significant or substantively important positive or negative effect. This results in a rating of no discernible effects, with a small extent of evidence. The supplemental findings include comparisons on social-emotional subscales and subgroups formed by mothers' race/ethnicity. The authors reported, and the WWC confirmed, a statistically significant and positive effect of *Head Start* on one social-emotional subscale for the 3-year-old cohort at the end of the intervention year. In addition, the authors reported, and the WWC confirmed, a statistically significant and positive effect of *Head Start* on two comparisons based on subgroups formed by mothers' race/ethnicity for the 3-year-old cohort at the end of the intervention year, including one composite measure and one subscale measure.

Table 3.3 Rating of effectiveness and extent of evidence for the social-emotional development domain

Rating of effectiveness	Criteria met
No discernible effects *No affirmative evidence of effects.*	In the one study that reported findings, the estimated impact of the intervention on outcomes in the *social-emotional development* domain was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Small	One study that included 3,693 3- and 4-year-old children in a nationally-representative sample reported evidence of effectiveness in the *social-emotional development* domain.

References

Study that meets WWC group design standards without reservations

U.S. Department of Health and Human Services, Administration for Children and Families. (2010). *Head Start impact study*. *Final report*. Washington, DC: Author. http://files.eric.ed.gov/fulltext/ED507845.pdf.

Additional sources:

Bitler, M. P., Hoynes, H. W., & Domina, T. (2013). *Experimental evidence on distributional effects of Head Start* (NBER Working Paper 20434). Cambridge, MA: National Bureau of Economic Research.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., & Downer, J. (2012). *Third grade follow-up to the Head Start impact study final report*. Washington, DC: U.S. Department of Health and Human Services, Office of Planning, Research and Evaluation. http://files.eric.ed.gov/fulltext/ED539264.pdf.

U.S. Department of Health and Human Services, Administration for Children and Families. (2005). *Head Start impact study: First year findings*. Washington, DC: Author. http://files.eric.ed.gov/fulltext/ED543015.pdf.

Studies that meet WWC group design standards with reservations

None.

Studies that do not meet WWC group design standards

Abbott-Shim, M., Lambert, R., & McCarty, F. (2003). A comparison of school readiness outcomes for children randomly assigned to a Head Start program and the program's wait list. *Journal of Education for Students Placed at Risk*, *8*(2), 191–214. The study does not meet WWC group design standards because it is a randomized controlled trial in which the combination of overall and differential attrition rates exceeds WWC standards for this area, and the subsequent analytic intervention and comparison groups are not shown to be equivalent.

Aughinbaugh, A. (2001). Does Head Start yield long-term benefits? *Journal of Human Resources*, *36*(4), 641–665. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Bassok, D. (2010). *Three essays on early childhood education policy* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3364495) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Bernardy, P. M. (2012). *Head Start: Assessing common explanations for the apparent disappearance of initial positive effects* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3547229) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Bumgarner, E. (2013). *Latino American children and school readiness: The role of early care arrangements and caregiver language* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3553548) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Chapman, M. A. (2012). *Impact of Head Start programs on reading skills of kindergarten students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3489478) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Chau, Y. H. (2005). *Investigating pre-K child care characteristics and Head Start on kindergarten outcomes: Analyses using the Early Childhood Longitudinal Study (ECLS-K)* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3147597) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Copple, C. E. (1987). *Path to the future: Long-term effects of Head Start in the Philadelphia school district*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families. http://files.eric.ed.gov/fulltext/ED289598.pdf. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Currie, J., & Thomas, D. (1993). Does Head Start make a difference? *The American Economic Review, 85*(3), 341–364. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Currie, J., & Thomas, D. (1998) *School quality and the longer-term effects of Head Start*. Washington, DC: National Bureau of Economic Research. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Currie, J., Thomas, D., & RAND Corp. (1996). *Does Head Start help Hispanic children? Labor and population program, working paper series 96-17*. Santa Monica, CA: RAND Corp. http://files.eric.ed.gov/fulltext/ED404008.pdf. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Deming, D. J. (2011). *Long-term impacts of educational interventions* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3414667) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Additional source:

Deming, D. J. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1*(3), 111–134.

Esteban, M. D. (1987). *A comparison of Head Start and non-Head Start reading readiness scores of low-income kindergarten children of Guam* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 8808677) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Fram, M. S., Kim, J., & Sinha, S. (2012). Early care and prekindergarten care as influences on school readiness. *Journal of Family Issues, 33*(4), 478–505. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Hebbeler, K. M. (1985). *Follow-up study of three cohorts of Head Start graduates*. Rockville, MD: Montgomery County Public Schools. http://files.eric.ed.gov/fulltext/ED263977.pdf. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Henry, G. T., Gordon, C. S., & Rickman, D. K. (2006). Early education policy alternatives: Comparing quality and outcomes of Head Start and state prekindergarten. *Educational Evaluation and Policy Analysis, 28*(1), 77–99. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Hickmon-Jenkins, C. G. (2000). *The effects of the Head Start experience upon reading readiness* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3023660) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Hunt, E. W. (1987). *A comparison of the academic achievement of urban second grade pupils with different forms of public preschool experience* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9007089) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention—there was only one unit assigned to one or both conditions.

Imai, K. (2003). *Evaluating early childhood interventions: Lessons from Head Start* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3104576) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Joo, M. (2006). *Long-term impacts of early childhood care and education on children's academic, behavior, and school outcomes: Is Head Start more effective than private preschools and no preschools for poor children*? (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3250506) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Kim, Y. (2008). *Educational achievement: The role of siblings, Head Start, and Catholic schools* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3314216) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Lee, V. E., Brooks-Gunn, J., & Schnur, E. (1988). Does Head Start work? A 1-year follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Developmental Psychology, 24*(2), 210–222. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Additional source:

Lee, V. E., Brooks-Gunn, J., Schnur, E., & Liaw, F. R. (1990). Are Head Start effects sustained? A longitudinal follow-up comparison of disadvantaged children attending Head Start, no preschool, and other preschool programs. *Child Development, 61*(2), 495–507.

Li, S. (2009). *Short-term and long-term effects of Head Start: A revisit using the Early Childhood Longitudinal Study* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3339353) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Lipscomb, S. T., Pratt, M. E., Schmitt, S. A., Pears, K. C., & Kim, H. K. (2013). School readiness in children living in non-parental care: Impacts of Head Start. *Journal of Applied Developmental Psychology, 34*(1), 28–37. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Lowenstein, A. E. (2009). *Fostering the socio-emotional adjustment of low-income children: The effects of universal pre-kindergarten and Head Start in Oklahoma* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3371816) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Marcon, R. A. (1996, March). *Head Start graduates: Making the transition from the early to the later childhood grades*. Paper presented at the Biennial Conference on Human Development, Birmingham, AL. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Mashburn, A. J. (2008). Quality of social and physical environments in preschools and children's development of academic, language, and literacy skills. *Applied Developmental Science, 12*(3), 113–127. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

McCoy, R. B. (1994). *A Head Start/public school blended preschool model: An early intervention study* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9429423) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Moore, T. J. (2013). *Effects of preschools on the academic outcome of children from low-income homes* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3507435) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Munoz, M. A., & Jefferson County Public Schools. (2001). *The critical years of education for at-risk students: The impact of an early childhood program on student learning*. Louisville, KY: Jefferson County Public Schools. http://files.eric.ed.gov/fulltext/ED456913.pdf. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Nystrom, P. J. (1988). *A longitudinal study to determine the effects of Head Start participation on reading achievement in grades kindergarten through six in Troy Public Schools* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 8910359) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention—there was only one unit assigned to one or both conditions.

Pigott, T. D., & Israel, M. S. (2005). Head Start children's transition to kindergarten: Evidence from the Early Childhood Longitudinal Study. *Journal of Early Childhood Research, 3*(1), 77–104. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Roy, A. (2003). *Evaluation of the Head Start program: Additional evidence from the NLSCM79 data* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3083526) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Sparagana, J. R. (2007). *The effect of a half-day pre-kindergarten program on school readiness in kindergarten* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3255649) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Stinson, J. E. (2012). *The impact of pre-K programs on student achievement and instructional leadership in rural Mississippi school districts* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3461316) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Turnage, A. L. (1998). *An examination of language outcomes for preschool-age children enrolled in different prekindergarten programs* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9928752) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention—there was only one unit assigned to one or both conditions.

Vaughn, B. E., Vollenweider, M., Bost, K. K., Azria-Evans, M., & Snider, J. B. (2003). Negative interactions and social competence for preschool children in two samples: Reconsidering the interpretation of aggressive behavior for young children. *Merrill-Palmer Quarterly, 49*(3), 245–278. The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Whittenberg, J. D. (2013). *Brigance, reading scores, and student preschool participation: Predictors of future academic achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3612233) The study does not meet WWC group design standards because it uses a quasi-experimental design in which the analytic intervention and comparison groups are not shown to be equivalent.

Williams, M. (1988). *Early childhood educational intervention: An analysis of Nicholas County, Kentucky Head Start program impacts from 1974–1986* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 8814371) The study does not meet WWC group design standards because the measures of effectiveness cannot be attributed solely to the intervention—there was only one unit assigned to one or both conditions.

Studies that are ineligible for review using the Early Childhood Education Evidence Review Protocol

Anderson, K., Foster, J., & Frisvold, D. (2010). Investing in health: The long-term impact of Head Start on smoking. *Economic Inquiry, 48*(3), 587–602. The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

Anderson, L. M., Shinn, C., Fullilove, M. T., Scrimshaw, S. C., Fielding, J. E., Normand, J., … U.S. Task Force on Community Preventive Services. (2003). The effectiveness of early childhood development programs: A systematic review. *American Journal of Preventive Medicine, 24*(Suppl3), 32–46. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Barnett, W. S., & Hustedt, J. T. (2005). Head Start's lasting benefits. *Infants & Young Children: An Interdisciplinary Journal of Special Care Practices, 18*(1), 16–24. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Bernstein, S. (2013). *Child care choice: Parental processes and consequences for research* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3527516) The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Carneiro, P., & Ginja, R. (2014). Long-term impacts of compensatory preschool on health and behavior: Evidence from Head Start. *American Economic Journal: Economic Policy 6*(4), 135–173. The study is ineligible for review because it does not examine an intervention implemented in a way that falls within the scope of the review.

Chambers, B., Cheung, A., Slavin, R. E., Smith, D., & Laurenzano, M. (2010). *Effective early childhood education programs: A systematic review*. Baltimore, MD: Johns Hopkins University, Center for Research and Reform in Education. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Cole, O. J., & Washington, V. (1986). A critical analysis of the assessment of the effects of Head Start on minority children. *Journal of Negro Education, 55*(1), 91–106. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Crahay, M. (1991). *Childcare and preschool effects: A review of Anglo-Saxon evaluative studies related to compensatory education and preschool education*. Liège, Belgium: University of Liège. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Educational Research Service. (1995). *Head Start*. Arlington, VA: Author. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Felix, C., & Frisvold, D. (2010). Health outcomes from Head Start participation. In D. Slottje & R. Tchernis (Eds.), *Current issues in health economics: Contributions to economic analysis* (pp. 115–138). Bradford, UK: Emerald Group Publishing. The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

Gamble, T. J., & Zigler, E. (1989). The Head Start synthesis project: A critique. *Journal of Applied Developmental Psychology, 10*(2), 267–274. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Garces, E., Currie, J., & Thomas, D. (2002). Longer term effects of Head Start. *The American Economic Review, 92*(4), 999–1012. The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

Additional source:

Garces, E., Thomas, D., & Currie, J. (2000). *Longer term effects of Head Start* (Unpublished manuscript). Department of Economics, University of California at Los Angeles.

Gibbs, C., Ludwig, J., & Miller, D. L. (2011). *Does Head Start do any lasting good?* (NBER Working Paper 17452). Cambridge, MA: National Bureau of Economic Research. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Gilliam, W. S., & Zigler, E. F. (2001). A critical meta-analysis of all evaluations of state-funded preschool from 1977 to 1998: Implications for policy, service delivery and program evaluation. *Early Childhood Research Quarterly, 15*(4), 441–473. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Gorey, K. M. (2001). Early childhood education: A meta-analytic affirmation of the short- and long-term benefits of educational opportunity. *School Psychology Quarterly, 16*(1), 9–30. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Greenfader, C. M., & Miller, E. B. (2014). The role of access to Head Start and quality ratings for Spanish-speaking dual language learners' (DLLs) participation in early childhood education. *Early Childhood Research Quarterly, 29*(3), 378–388. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Haas, L. E. (2011). *Formal and informal measures of reading and math achievement as a function of early childhood program participation among kindergarten through eighth grade students* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3484769) The study is ineligible for review because it does not use a comparison group design or a single-case design.

Hale, B., Seitz, V., & Zigler, E. (1990). Health services and Head Start: A forgotten formula. *Journal of Applied Developmental Psychology, 11*(4), 447–458. The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

Herman, A. D., & Mayer, G. G. (2004). Reducing the use of emergency medical resources among Head Start families: A pilot study. *Journal of Community Health: The Publication for Health Promotion and Disease Prevention, 29*(3), 197–208. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Isaacs, J. B. (2008). *Impacts of early childhood programs*. Washington, DC: The Brookings Institution and First Focus. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Jessup, P. A. (2008). Learning research: Insights from Head Start. *Journal of Early Childhood Research, 6*(1), 51–57. The study is ineligible for review because it does not use a comparison group or single case design.

Lee, V. E., & Loeb, S. (1995). Where do Head Start attendees end up? One reason why preschool effects fade out. *Educational Evaluation and Policy Analysis, 17*(1), 62–82. The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

Li, W. (2014). *Center-based early childhood education: Curriculum, implementation, and intensity* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3564827) The study is ineligible for review because it does not use a comparison group design or a single-case design.

Love, J. M., Chazan-Cohen, R., & Raikes, H. (2007). Forty years of research knowledge and use: From Head Start to Early Head Start and beyond. In J. L. Aber, S. J. Bishop-Josef, S. M. Jones, K. T. McLearn, & D. A. Phillips (Eds.), *Child development and social policy: Knowledge for action* (pp. 79–95). Washington, DC: American Psychological Association. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Love, J. M., Grover, J., & RMC Research Corp. (1987). *Study of Head Start recruitment and enrollment: Final report*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children, Youth and Families. http://files.eric.ed.gov/fulltext/ED283607.pdf. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Ludwig, J., & Phillips, D. (2008). The long-term effects of Head Start on low-income children. *Annals of the New York Academy of Sciences, 40*, 1–12. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Ludwig, J., & Phillips, D. A. (2007). *The benefits and costs of Head Start* (NBER Working Paper 12973). Cambridge, MA: National Bureau of Economic Research. http://files.eric.ed.gov/fulltext/ED521701.pdf. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Lumeng, J., Kaciroti, N., & Frisvold, D. (2010). Changes in body mass index Z score over the course of the academic year among children attending Head Start. *Academic Pediatrics, 10*(3), 179–186. The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

McGroder, S. M. (1990). *Head Start: What do we know about what works*. Washington, DC: U.S. Dept. of Health and Human Services. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

McKey, R. H. (1985). *The impact of Head Start on children, families and communities. Final report of the Head Start evaluation, synthesis and utilization project*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children, Youth and Families. http://files.eric.ed.gov/fulltext/ED263984.pdf. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Mervis, J. (2011). Giving children a Head Start is possible—but it's not easy. *Science, 333*(6045), 956–957. The study is ineligible for review because it does not use a comparison group design or a single-case design.

Nielsen, W. L. (1989). The longitudinal effects of project head start on students' overall academic success: A review of the literature. *International Journal of Early Childhood, 21*(1), 35–42. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Research-based responses to key questions about the 2010 Head Start impact study. (2011). *Child Trends: Early Childhood Highlights, 2*(1). The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Reynolds, A. J. (Ed.). (2010). *Childhood programs and practices in the first decade of life: A human capital integration*. New York: Cambridge University Press. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Schweinhart, L. J. (2003). The three types of early childhood programs in the United States. In A. J. Reynolds (Ed.), *Early childhood programs for a new century* (pp. 241–254). Washington, DC: Child Welfare League of America, Inc. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Schweinhart, L. J., & ERIC Clearinghouse on Elementary and Early Childhood Education. (2001). *Recent evidence on preschool programs. ERIC digest*. Champaign IL: ERIC Clearinghouse on Elementary and Early Childhood Education. http://files.eric.ed.gov/fulltext/ED458046.pdf. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Schweinhart, L. J., & Weikart, D. P. (1986). What do we know so far? A review of the Head Start synthesis project. *Young Children, 41*(2), 49–55. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Shager, H. M., Schindler, H. S., Magnuson, K. A., Duncan, G. J., Yoshikawa, H., & Hart, C. M. D. (2013). Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes. *Educational Evaluation and Policy Analysis, 35(*1), 76–95. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Shimoni, R. (1990). *A historical overview of the development of early childhood services*. http://files.eric.ed.gov/fulltext/ED334000.pdf. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Sprigle, J. E., & Schaefer, L. (1985). Longitudinal evaluation of the effects of two compensatory preschool programs on fourth- through sixth-grade students. *Developmental Psychology, 21*(4), 702. The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

Swadener, B. B., Dunlap, S. K., & Nespeca, S. M. (1995). Family literacy and social policy: Parent perspectives and policy implications. *Reading & Writing Quarterly, 11*(3), 267–283. The study is ineligible for review because it does not use a comparison group design or a single-case design.

U.S. Department of Health and Human Services, Administration for Children, Youth and Families, Head Start Bureau. (1993). *Head Start: A child development program*. Washington, DC: Author. The study is ineligible for review because it does not use a comparison group design or a single-case design.

U.S. Department of Health and Human Services, Office of Inspector General. (1993). *Evaluating Head Start expansion through performance indicators*. OEI-09-91-00762. Washington, DC: U.S. Department of Health and Human Services. The study is ineligible for review because it does not use a comparison group design or a single-case design.

U.S. General Accounting Office (2000). *Early childhood programs: Characteristics affect the availability of school readiness information*. GAO/HEHS-00-38. Washington, DC: Author. Retrieved from http://www.gao.gov/assets/230/228763.pdf. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Whiteside-Mansell, L., Bradley, R., McKelvey, L., & Lopez, M. (2009). Center-based Early Head Start and children exposed to family conflict. *Early Education and Development*, 20(6), 942–957. The study is ineligible for review because it does not include an outcome within a domain specified in the protocol.

Zeece, P. D., & Wang, A. (1998). Effects of the family empowerment and transitioning program on child and family outcomes. *Child Study Journal, 28*(3), 161–178. The study is ineligible for review because it does not examine an intervention implemented in a way that falls within the scope of the review—the intervention is bundled with other components.

Zhai, F., Waldfogel, J., & Brooks-Gunn, J. (2013). Head Start, prekindergarten, and academic school readiness: A comparison among regions in the United States. *Journal of Social Service Research, 39*(3), 345–364. The study is ineligible for review because it does not examine an intervention implemented in a way that falls within the scope of the review.

Zigler, E. F., & Styfco, S. J. (1996). Head Start and early childhood intervention: The changing course of social science and social policy. In E. F. Zigler, S. L. Kagan, & N. W. Hall (Eds.), *Children, families, and government: Preparing for the twenty-first century* (pp. 132–155). New York: Cambridge University Press. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Zigler, E. F., & Styfco, S. J. (2003). The federal commitment to preschool education: Lessons from and for Head Start. In A. J. Reynolds (Ed.), *Early childhood programs for a new century* (pp. 3–33). Washington, DC: Child Welfare League of America, Inc. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Zigler, E. F., & Styfco, S. J. (2004). *The Head Start debates*. Baltimore, MD: Brookes Publishing Company. The study is ineligible for review because it is a secondary analysis of the effectiveness of an intervention, such as a meta-analysis or research literature review.

Appendix A: Research details for DHHS ACF (2010)

U.S. Department of Health and Human Services, Administration for Children and Families. (2010). *Head Start impact study. Final report*. Washington, DC. http://files.eric.ed.gov/fulltext/ED507845.pdf.[20]

Additional sources:

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., & Downer, J. (2012). *Third grade follow-up to the Head Start impact study final report*. Washington, DC: U.S. Department of Health and Human Services Office of Planning, Research and Evaluation. http://files.eric.ed.gov/fulltext/ED539264.pdf.

U.S. Department of Health and Human Services, Administration for Children and Families. (2005). *Head Start impact study: First year findings*. Washington, DC. http://files.eric.ed.gov/fulltext/ED543015.pdf.

Table A. Summary of findings **Meets WWC group design standards without reservations**

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
General reading achievement	3,697 children	+13	Yes
Mathematics achievement	1,617 children	+3	No
Social-emotional development	3,693 children	+1	No

Setting

The study was conducted using a nationally-representative sample of *Head Start* programs in the United States.[21]

Head Start programs in Puerto Rico were included in the original sample, but were analyzed separately and are not included in this report, since the assessments were provided only in Spanish, which is outside the scope of the review protocol for the Early Childhood Education topic area (version 3.0).

Study sample

The sample was created using a multistage process. All *Head Start* and delegate agencies in fiscal year 1998–99 were stratified by geography and demographic characteristics, and a random sample was then drawn from this list.[22] *Head Start* and delegate agencies were excluded from the sample pool if they were new grantees, were administered by American Indian/Alaska Native tribal organizations, had participated in the *Head Start* Family and Child Experiences Survey (FACES) 2000,[23] ran programs that were exclusively *Early Head Start* or *Migrant and Seasonal Head Start*, or had operated in communities in which most children participated in *Head Start* (this is because these programs were "saturated," meaning there would be a limited chance of forming a comparison group, since so many children were already being served).

Subsequently, eligible *Head Start* programs were randomly sampled from within delegate agencies. Similar to the criteria used above to exclude *Head Start* and delegate agencies, *Head Start* programs were excluded from the sample pool if they were saturated, had closed or merged with another program, were co-operated with a non-*Head Start* agency (e.g., a private preschool program), or were exclusively *Early Head Start* or *Migrant and Seasonal Head Start* programs.

Finally, children were randomly selected from the applicant pool of each program and then randomly assigned either to be offered *Head Start* or to be in the comparison group. Programs were allowed to exclude a limited number of children from the random assignment process if they were thought to be "high-risk" and in particular need of *Head Start* services. These children were not included in the impact analysis.

The study design and presentation of findings focused on two cohorts of children:

Three-year-old cohort: Children who were 3-years-old when applying to *Head Start*. The baseline sample included 383 *Head Start* programs, 1,466 children in the *Head Start* group, and 988 children in the comparison group. The analytic sample (after attrition) included at most 2,062 children (1,278 intervention and 784 comparison) for the general reading achievement and social-emotional development outcome domains.

Four-year-old cohort: Children who were 4-years-old when applying to *Head Start*. The baseline sample included 383 *Head Start* programs, 1,192 children in the *Head Start* group, and 815 children in the comparison group. The analytic sample (after attrition) included at most 1,635 children (1,008 intervention and 627 comparison), depending on the outcome.

Children in both cohorts were followed through the spring of third grade. Approximately 50% of the 3-year-old children who were originally assigned to the comparison group enrolled in *Head Start* as 4-year-olds. As a result, the desired contest was not maintained after the first year. Any impacts examined with the 3-year-old cohort after the first year of *Head Start* were determined to not be a test of the effectiveness of *Head Start* and are not included in this intervention report.

Intervention group

Head Start includes diverse program models, and the study intervention group did as well. The intervention group included: center-based programs with home visits (the most common type), programs in which *Head Start* staff visited families at their homes, family child care programs, and programs that combined these models.

Individual *Head Start* programs maintained their standard practices during the study. The programs in the *Head Start* group varied in terms of their quality, the specific types of services provided, and the numbers of months and hours the programs were available. In addition, children's attendance levels varied.

Comparison group

Parents of children in the comparison group were free to enroll their children in any program other than the *Head Start* programs in the study (or to not enroll them in any program). Consequently, children in the comparison group experienced diverse types of early care and education settings ranging from parent-only care to programs that were similar in type and services to *Head Start*.

Authors reported that 17.3% of the 3-year-old baseline comparison group were enrolled in *Head Start* programs that were not part of the study in spite of their study group assignment.[24] Authors reported that 13.9% of the 4-year-old baseline comparison group were enrolled in *Head Start* programs that were not part of the study in spite of their study group assignment.[25]

Outcomes and measurement

The following outcome measures were administered to both the 3-year-old and 4-year-old cohorts at the end of the intervention year for contrasts that meet WWC group design standards without reservations. In the general reading achievement domain, the authors used the PELS. In the social-emotional development domain, the authors used the Total Problem Behavior Scale, the Social Competencies Checklist, and the Social Skills and Positive Approaches to Learning Scale. In addition, in the mathematics achievement domain, the Counting Bears Test was used and meets WWC group design standards for only the 4-year-old cohort. For a more detailed description of these outcome measures, see Appendix B.

Follow-up findings from kindergarten and grade 1 are presented as supplemental findings for the 4-year-old cohort in the mathematics achievement, social-emotional development, alphabetics, cognition, comprehension, and language development domains. Follow-up findings for the 4-year-old cohort in grade 3 are presented as supplemental findings in the general reading achievement, mathematics achievement, social-emotional development, alphabetics, and comprehension domains. Subscale findings for the 3- and 4-year old cohorts at the end of the intervention year are presented as supplemental findings in the social-emotional development domain; follow-up subscale findings from this domain are also presented for the 4-year-old cohort from kindergarten, grade 1, and grade 3. Findings based on mothers' race/ethnicity in the mathematics achievement, social-emotional development, alphabetics, cognition, comprehension, and language development domains are also presented as supplemental findings. None of the supplemental findings factor into the intervention's ratings of effectiveness.

Some outcomes are ineligible for review under the Early Childhood Education topic area and were excluded from this review, including the Early Childhood Environment Rating Scale—Revised, the Arnett Scale of Lead Teacher Behavior, the type of curricula used, the type of child care arrangement, parent interview questions related to health and parenting practices, and parent and teacher reports about their relationship with the child, grade promotion, and school accomplishments. The review prioritized direct child assessments over teacher reports and also prioritized standardized measures over non-standardized measures. When direct assessments were not available, the review prioritized adult-reported measures over child reports, and thus excluded child-reported measures of internalizing and externalizing problems, peer relations, and school performance.

Many of the follow-up, subscale, and race/ethnicity subgroup comparisons do not meet WWC group design standards, either because the outcomes do not meet review requirements or because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated. These comparisons are not included in Appendix D.

Support for implementation

The study did not report information on the support or professional development offered in *Head Start* programs.

Appendix B: Outcome measures for each domain

General reading achievement	
*Early Childhood Longitudinal Study–Kindergarten Reading (ECLS-K Reading)**	The ECLS-K is a standardized measure of children's reading skills including comprehension, decoding, and vocabulary (as cited in DHHS ACF, 2012).
Parent Emergent Literacy Scale (PELS)	The PELS measures children's skills in five areas using parent ratings. The areas include letter recognition, counting, name writing (real or pretend), and primary color identification. The instrument was developed for the FACES 2000 study (as cited in DHHS ACF, 2010).
Mathematics achievement	
Counting Bears Test	The Counting Bears Test measures children's counting abilities and their understanding of one-to-one correspondence. This task was adapted for the FACES study (U.S. Department of Health and Human Services, 2001) from the Comprehensive Assessment Program (CAP) Early Childhood Diagnostic Instrument (Mason & Stewart, 1989; as cited in DHHS ACF, 2010).
*McCarthy Scales of Children's Abilities-Draw-A-Design Task**	The Draw-A-Design task is a standardized measure that assesses children's perceptual motor and pre-writing skills (McCarthy, 1970, 1972; as cited in DHHS ACF, 2010).
Woodcock-Johnson III Tests of Achievement (WJ-III), Applied Problems Subtest[a]*	The WJ-III, Applied Problems Subtest is a standardized measure of children's abilities to solve mathematical problems presented orally (Woodcock et al., 2001; as cited in DHHS ACF, 2010).
WJ-III, Calculation Subtest[a]*	The WJ-III, Calculation Subtest is a standardized measure of children's knowledge of numbers and abilities to perform calculations (Woodcock et al., 2001; as cited in DHHS ACF, 2010).
WJ-III, Math Reasoning Test[a]*	The WJ-III, Math Reasoning Test is a composite of two subtests (Applied Problems and Quantitative Concepts). It is a standardized measure of children's mathematical knowledge and reasoning (Woodcock et al., 2001; as cited in DHHS ACF, 2010).
WJ-III, Quantitative Concepts Test[a]*	The WJ-III, Quantitative Concepts Test is a composite of two subtests (Concepts and Number Series). It is a standardized measure of children's abilities to count, identify shapes, patterns, numbers, and series, as well as their knowledge of mathematical concepts and terms (Woodcock, McGrew, & Mather, 2001; as cited in DHHS ACF, 2010).
Social-emotional development	
*Adjustment Scales for Preschool Intervention (ASPI), Aggressive Behavior Dimension**	The ASPI, Aggressive Behavior Dimension measures 22 items associated with aggressive behaviors based on teacher report. The teacher is asked to choose behaviors (out of 144) that were demonstrated by the child during specific types of classroom situations over the previous 2 months. The score is the number of behaviors indicated in each dimension. It was based on the Adjustment Scales for Children and Adolescents (ASCA) (Lutz, Fantuzzo, & McDermott, 2000; as cited in DHHS ACF, 2010).
*ASPI, Inattentive/Hyperactive Dimension**	The ASPI, Inattentive/Hyperactive Dimension measures the extent to which children demonstrate 10 behaviors associated with inattention, impulsivity, or hyperactivity (Lutz et al., 2000; as cited in DHHS ACF, 2010).
*ASPI, Oppositional Dimension**	The ASPI, Oppositional Dimension measures the extent to which children demonstrate 11 behaviors associated with moodiness and controlling behaviors (Lutz et al., 2000; as cited in DHHS ACF, 2010).
*ASPI, Problems with Peer Interaction Dimension**	The ASPI, Problems with Peer Interaction Dimension measures the extent to which children demonstrate 24 problem behaviors over the course of six types of peer situations (Lutz et al., 2000; as cited in DHHS ACF, 2010).
*ASPI, Problems with Structured Learning Dimension**	The ASPI, Problems with Structured Learning Dimension measures the extent to which children demonstrate 40 problem behaviors over the course of seven types of structured classroom situations (Lutz et al., 2000; as cited in DHHS ACF, 2010).
*ASPI, Problems with Teacher Interaction Dimension**	The ASPI, Problems with Teacher Interaction Dimension measures the extent to which children demonstrate 30 problem behaviors over the course of six types of classroom situations that include teachers (Lutz et al., 2000; as cited in DHHS ACF, 2010).
*ASPI, Shy/Socially Reticent Dimension**	The ASPI, Socially Reticent Dimension measures the extent to which children demonstrate 12 behaviors associated with shyness and hesitancy (Lutz et al., 2000; as cited in DHHS ACF, 2010).

*ASPI, Withdrawn-Low Energy Behavior Dimension**	The ASPI, Withdrawn-Low Energy Dimension measures the extent to which children demonstrate 18 behaviors associated with lack of energy and activity (Lutz et al., 2000; as cited in DHHS ACF, 2010).
Social Competencies Checklist	The Social Competencies Checklist measures children's social skills based on parents' reports of the extent to which children exhibit 12 behaviors or characteristics (e.g., "takes care of personal belongings"); (Developing Skills Checklist, 1990; as cited in DHHS ACF, 2010).
Social Skills and Positive Approaches to Learning Scale	The Social Skills and Positive Approaches to Learning Scale measures children's social skills (e.g., cooperative and empathic behavior) and approaches to learning (e.g., curiosity, imagination, openness to challenges, and positive attitudes about gaining skills and knowledge). It is a researcher-developed composite of parents' ratings on seven items. The measure is based on one used in the FACES study (U.S. Department of Health and Human Services, 2001) that was in turn based on a modified Achenbach Classroom Behavior Checklist (Achenbach, Edelbrock, & Howell, 1987; as cited in DHHS ACF, 2010).
Total Problem Behavior Scale (TPBS)	The Total Problem Behavior Scale measures the extent to which children exhibit problem behaviors. It is a composite created from parents' ratings that combines items in three subscales measuring: aggressive or defiant behavior; inattentive or hyperactive behavior; and shy, withdrawn, or depressed behavior. Parents are asked to judge whether the behavioral description offered in each of 14 items is "not true," "sometimes true," or "very true" of the child (Achenbach et al., 1987; as cited in DHHS ACF, 2010).
*TPBS, Aggressive Behavior Subscale**	The Aggressive Behavior Subscale of the Total Problem Behavior Scale measures the extent to which children exhibit aggressive or defiant behavior. It is a composite created from parents' ratings of children's behavior on four items (Achenbach et al., 1987; as cited in DHHS ACF, 2010).
*TPBS, Hyperactive Behavior Subscale**	The Hyperactive Behavior Subscale of the Total Problem Behavior Scale measures the extent to which children exhibit hyperactive or inattentive behavior. It is a composite created from parents' ratings of children's behavior on three items (Achenbach et al., 1987; as cited in DHHS ACF, 2010).
*TPBS, Withdrawn Behavior Subscale**[b]	The Withdrawn Behavior Subscale of the Total Problem Behavior Scale measures the extent to which children exhibit shy, withdrawn, or depressed behavior. It is a composite created from parents' ratings of children's behavior on three items (Achenbach et al., 1987; as cited in DHHS ACF, 2010).
Alphabetics	
Letter identification construct	
*Letter Naming Task**	The Letter Naming Task is a measure of children's abilities to recognize letters of the alphabet. The task was modified for use in the FACES study (DHHS, 2001) from a test used in the Head Start Quality Research Center's curricular intervention studies. The task was administered in English to bilingual children, but responses in English or Spanish were accepted (as cited in DHHS ACF, 2010).
Phonological awareness construct	
*Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP), Elision Subtest**	The Pre-CTOPPP, Elision Subtest is a standardized measure of children's ability to identify and manipulate sounds in spoken words (Lonigan, Wagner, Torgesen, & Rashotte, 2002; as cited in DHHS ACF, 2010).
Phonics construct	
*WJ-III, Basic Reading Skills Test**[a]	The WJ-III, Basic Reading Skills Test is a composite of two subtests, Letter-Word Identification and Word Attack. It is a standardized measure of children's abilities to use phonics, recognize words by sight, and use structural analysis (Woodcock et al., 2001; as cited in DHHS ACF, 2010).
*WJ-III, Letter-Word Identification Subtest**[a]	The WJ-III, Letter-Word Identification Subtest is a standardized measure of children's abilities to identify letters and words in English (Woodcock et al., 2001; as cited in DHHS ACF, 2010).
*WJ-III, Spelling Subtest**[a]	The WJ-III, Spelling Subtest is a standardized measure of children's abilities to write letters and words presented orally in English as well as pre-writing skills such as line-drawing and letter copying (Woodcock et al., 2001; as cited in DHHS ACF, 2010).
*WJ-III, Word Attack Subtest**[a]	The WJ-III, Word Attack Subtest is a standardized measure of children's abilities to produce the sounds associated with letters and to read real and nonsense words aloud using phonics and structural analysis (Woodcock et al., 2001; as cited in DHHS ACF, 2010).

*WJ-III Writing Sample Subtest**	The WJ-III, Writing Sample Subtest is a standardized measure of children's ability to provide written responses to various demands, such as completing sentences, passages, or describing pictures (Woodcock et al., 2001; as cited in DHHS, ACF, 2010).
Cognition	
*Color Naming/Identification Task**	The Color Naming/Identification Task measures children's color recognition and naming abilities. In the assessment, children are asked to point to and name 10 colors. This task was adapted for the FACES study (DHSS, 2001) from the Comprehensive Assessment Program Early Childhood Diagnostic Instrument (Mason & Stewart, 1989; as cited in DHHS ACF, 2010).
*WJ-III, Academic Applications**[a]	The WJ-III, Academic Applications is a composite of three subtests: Passage Comprehension, Applied Problems, and Writing Samples. It is a standardized measure of children's use of academic skills (as cited in DHHS ACF, 2010).
*WJ-III, Academic Skills**[a]	The WJ-III, Academic Skills is a composite of three subtests: Letter-Word Identification, Spelling, and Calculation. It is a standardized measure of children's basic academic skills in the areas of decoding, mathematical calculations, and spelling (Woodcock et al., 2001; as cited in DHHS ACF, 2010).
*WJ-III, Pre-Academic Skills**[a]	The WJ-III, Pre-Academic Skills is a composite of three subtests: Letter-Word Identification, Spelling, and Applied Problems (as cited in DHHS ACF, 2010).
Comprehension	
Reading comprehension construct	
*WJ-III, Passage Comprehension Subtest**[a]	The WJ-III Passage Comprehension Subtest is a standardized measure of children's abilities to provide missing words in passages using contextual information provided in the passages. The assessment begins by testing the ability to match symbols of objects with pictures of them and progresses through passages with increasing levels of difficulty in terms of length, vocabulary, and semantic complexity (Woodcock et al., 2001; as cited in DHHS ACF, 2010).
Vocabulary development construct	
*Peabody Picture Vocabulary Test, Third Edition–Adapted (PPVT-Adapted)**	The PPVT-Adapted is a standardized measure of receptive vocabulary in which children point to the pictures of named objects and actions that represents their meanings. The study employed a shortened version of this assessment, developed using maximum likelihood Item Response Theory(Dunn, Dunn, & Dunn, 1997; as cited in DHHS ACF, 2010).
Language development	
*WJ-III, Oral Comprehension Subtest**[a]	The WJ-III Oral Comprehension Subtest is a standardized measure of children's abilities to comprehend a short orally-presented passage. In the assessment, children are asked to fill in missing words using syntactic and semantic information in the passage(Woodcock et al., 2001; as cited in DHHS ACF, 2010).

* This outcome is considered a supplemental outcome because: (1) it was either a subtest, subscale, or an outcome that was measured after the intervention year, so it does not represent an immediate impact of the intervention or (2) the main analyses for the outcome did not meet WWC group design standards but the subgroup analyses did meet standards. For this reason, contrasts for these outcomes that meet WWC group design standards are presented in Appendix D, but do not contribute to the rating of effectiveness or the extent of evidence.

[a] The rule for ending the administration of all WJ-III subscales was changed from the standard ceiling rule (stop the test after six incorrect answers have been given) to a lower ceiling rule (stop after three incorrect responses). The change was made so that the results from the current study would be comparable to those of the FACES study (U.S. Department of Health and Human Services, 2001; as cited in DHHS ACF, 2010), which used the adapted rules to reduce the test burden on the young children being tested. Standard ceiling rules were used for the first-grade test administration.

[b] The *TPBS, Withdrawn Behavior Subscale* for the 4-year-old cohort was only eligible for review during first grade (in 2005) because for the Head Start and Kindergarten years (2003 and 2004 respectively) the reported measure of internal consistency was below the reliability threshold set forth by the Early Childhood Education review protocol (version 3.0). The internal consistency of this measure for the 4-year-old cohort during third grade was not provided by the authors and cannot be assumed to have met the threshold.

Appendix C.1: Findings included in the rating for the general reading achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF, 2010[a]								
Parent Emergent Literacy Scale (PELS)	3-year-olds, Head Start year	2,062 children	2.86 (1.48)	2.35 (1.38)	0.51	0.35	+14	< .01
PELS	4-year-olds, Head Start year	1,635 children	3.76 (1.35)	3.35 (1.40)	0.41	0.30	+12	< .01
Domain average for general reading achievement (DHHS ACF, 2010)						**0.33**	**+13**	**Statistically significant**
Domain average for general reading achievement across all studies						**0.33**	**+13**	**na**

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. The WWC-computed average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. The statistical significance of each study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable.

[a] For DHHS ACF (2010), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. This study is characterized as having a statistically significant positive effect because the effect for at least one measure within the domain is positive and statistically significant, and no effects are negative and statistically significant, accounting for multiple comparisons. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix C.2: Findings included in the rating for the mathematics achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF, 2010[a]								
Counting Bears Test	4-year-olds, Head Start year	1,617 children	0.59 (0.49)	0.55 (0.50)	0.04	0.08	+3	.19
Domain average for mathematics achievement (DHHS ACF, 2010)						**0.08**	**+3**	**Not statistically significant**
Domain average for mathematics achievement across all studies						**0.08**	**+3**	**na**

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. The statistical significance of each study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable.

[a] For DHHS ACF (2010), no corrections for clustering or multiple comparisons and no difference-in-differences adjustments were needed. The *p*-value presented here was reported in the original study. This study is characterized as having an indeterminate effect because the mean effect is neither statistically significant nor substantively important according to WWC criteria. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix C.3: Findings included in the rating for the social-emotional development domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF, 2010[a]								
Total Problem Behavior Scale (negative)	3-year-olds, Head Start year	2,062 children	5.80 (3.56)	6.24 (3.74)	0.44	0.12	+5	.05
Social Competencies Checklist	3-year-olds, Head Start year	2,061 children	10.95 (1.40)	10.99 (1.24)	–0.04	–0.03	–1	.54
Social Skills and Positive Approaches to Learning Scale	3-year-olds, Head Start year	2,062 children	12.41 (1.75)	12.38 (1.70)	0.03	0.02	+1	.74
Total Problem Behavior Scale (negative)	4-year-olds, Head Start year	1,629 children	5.60 (3.83)	5.80 (3.33)	0.20	0.05	+2	.41
Social Competencies Checklist	4-year-olds, Head Start year	1,631 children	11.01 (1.46)	11.06 (1.19)	–0.05	–0.04	–1	.67
Social Skills and Positive Approaches to Learning Scale	4-year-olds, Head Start year	1,629 children	12.46 (1.79)	12.48 (1.64)	–0.02	–0.01	0	.89
Domain average for social-emotional development (DHHS ACF, 2010)						**0.02**	**+1**	**Not statistically significant**
Domain average for social-emotional development across all studies						**0.02**	**+1**	**na**

Table Notes: For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The outcome measures followed by "(negative)" imply that a lower score is considered a desirable outcome, so the signs of all WWC calculated statistics (mean difference, effect size, and improvement index) were adjusted to reflect this. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. The WWC-computed average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. The statistical significance of each study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable.

[a] For DHHS ACF (2010), no corrections for clustering or multiple comparisons were needed. The *p*-values presented here were reported in the original study. The study-reported *p*-value for the total problem behavior scale for 3-year-olds was .053; this *p*-value is not statistically significant. This study is characterized as having an indeterminate effect because the mean effect is neither statistically significant nor substantively important, accounting for multiple comparisons. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix D.1: Description of supplemental findings for the general reading achievement domain, follow-up findings

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Puma et al., 2012[a] (4-year-old cohort)								
Early Childhood Longitudinal Study–Kindergarten (ECLS-K)	4-year-old cohort, Grade 3	1,414 students	98.61 (19.63)	96.63 (20.24)	1.98	0.10	+4	.14

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For Puma et al. (2012), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The p-values presented here were reported in the original study. Sample sizes and standard deviations were obtained in an author query.

Appendix D.2A: Description of supplemental findings for the mathematics achievement domain, follow-up findings

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (4-year-old cohort)								
Woodcock-Johnson III Tests of Achievement (WJ-III), Applied Problems Subtest	4-year-old cohort, Kindergarten	1,533 students	426.59 (19.55)	426.32 (21.85)	0.27	0.01	+1	.87
WJ-III, Math Reasoning Test	4-year-old cohort, Kindergarten	1,533 students	434.15 (16.53)	434.12 (17.68)	0.03	0.00	0	.98
WJ-III, Quantitative Concepts Test	4-year-old cohort, Kindergarten	1,530 students	441.83 (17.77)	441.88 (17.16)	–0.05	0.00	0	.97
WJ-III, Applied Problems Subtest	4-year-old cohort, Grade 1	1,526 students	455.16 (19.30)	454.13 (19.82)	1.03	0.05	+2	.41
WJ-III, Calculation Subtest	4-year-old cohort, Grade 1	1,519 students	461.76 (18.29)	460.46 (19.40)	1.30	0.07	+3	.25
WJ-III, Math Reasoning Test	4-year-old cohort, Grade 1	1,526 students	458.36 (17.18)	457.67 (17.48)	0.69	0.04	+2	.58
WJ-III, Quantitative Concepts Test	4-year-old cohort, Grade 1	1,524 students	461.79 (17.49)	461.28 (17.99)	0.51	0.03	+1	.71
Puma et al. (2012)[a] (4-year-old cohort)								
WJ-III, Applied Problems Subtest	4-year-old cohort, Grade 3	1,422 students	486.96 (20.37)	487.70 (19.40)	–0.74	–0.04	–1	.60
WJ-III, Calculation Subtest	4-year-old cohort, Grade 3	1,422 students	491.28 (15.75)	491.52 (16.35)	–0.24	–0.02	–1	.83

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) and Puma et al. (2012) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. For Puma et al. (2012), sample sizes and standard deviations were obtained in an author query.

Appendix D.2B: Description of supplemental findings for the mathematics achievement domain, based on mothers' race/ethnicity

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (3-year-old cohort)								
McCarthy Scales of Children's Abilities, Draw-A-Design Task	3-year-old cohort, White/Other, Head Start year	672	3.18 (1.19)	3.00 (1.15)	0.18	0.15	+6	.11
Counting Bears Test	3-year-old cohort, White/Other, Head Start year	659	0.33 (0.46)	0.31 (0.45)	0.02	0.04	+2	.70
WJ-III, Applied Problems Subtest	3-year-old cohort, White/Other, Head Start year	665	387.98 (27.44)	382.22 (29.20)	5.76	0.20	+8	.07
McCarthy Scales of Children's Abilities, Draw-A-Design Task	3-year-old cohort, Hispanic, Head Start year	654	3.33 (1.19)	3.19 (1.15)	0.14	0.12	+5	.25
DHHS ACF (2010)[b] (4-year-old cohort)								
McCarthy Scales of Children's Abilities, Draw-A-Design Task	4-year-old cohort, White/Other, Head Start year	594	4.47 (2.16)	4.23 (1.95)	0.24	0.11	+5	.25
Counting Bears Test	4-year-old cohort, White/Other, Head Start year	588	0.58 (0.49)	0.62 (0.50)	–0.04	–0.08	–3	.32
WJ-III, Applied Problems Subtest	4-year-old cohort, White/Other, Head Start year	593	406.34 (23.96)	404.44 (26.99)	1.90	0.08	+3	.43
WJ-III, Applied Problems Subtest	4-year-old cohort, White/Other, Kindergarten	560	431.39 (19.55)	432.35 (21.85)	–0.96	–0.05	–2	.53
WJ-III, Math Reasoning Test	4-year-old cohort, White/Other, Kindergarten	560	436.65 (16.53)	438.06 (17.68)	–1.41	–0.08	–3	.34
WJ-III, Quantitative Concepts Test	4-year-old cohort, White/Other, Kindergarten	558	441.89 (17.77)	443.83 (17.16)	–1.94	–0.11	–4	.27

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
WJ-III, Applied Problems Subtest	4-year-old cohort, White/ Other, Grade 1	548	459.75 (19.30)	458.16 (19.82)	1.59	0.08	+3	.38
WJ-III, Calculation Subtest	4-year-old cohort, White/ Other, Grade 1	546	460.99 (18.29)	460.19 (19.40)	0.80	0.04	+2	.71
WJ-III, Math Reasoning Test	4-year-old cohort, White/ Other, Grade 1	548	462.09 (17.18)	460.68 (17.48)	1.41	0.08	+3	.43
WJ-III, Quantitative Concepts Test	4-year-old cohort, White/ Other, Grade 1	548	464.56 (17.49)	463.28 (17.99)	1.28	0.07	+3	.50
Counting Bears Test	4-year-old cohort, Hispanic, Head Start year	672	0.57 (0.49)	0.49 (0.50)	0.08	0.16	+6	.09
McCarthy Scales of Children's Abilities, Draw-A-Design Task	4-year-old cohort, Hispanic, Head Start year	681	4.80 (2.16)	4.57 (1.95)	0.23	0.11	+4	.26
WJ-III, Applied Problems Subtest	4-year-old cohort, Hispanic, Head Start year	674	389.02 (23.96)	384.54 (26.99)	4.48	0.18	+7	.31
WJ-III, Applied Problems Subtest	4-year-old cohort, Hispanic, Grade 1	634	452.52 (19.30)	452.10 (19.82)	0.42	0.02	+1	.81
WJ-III, Calculation Subtest	4-year-old cohort, Hispanic, Grade 1	632	463.28 (18.29)	463.07 (19.40)	0.21	0.01	0	.90
WJ-III, Math Reasoning Test	4-year-old cohort, Hispanic, Grade 1	634	455.79 (17.18)	456.19 (17.48)	–0.40	–0.02	–1	.82
WJ-III, Quantitative Concepts Test	4-year-old cohort, Hispanic, Grade 1	633	459.44 (17.49)	460.33 (17.99)	–0.89	–0.05	–2	.65

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF, 2010 (3-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

[b] For DHHS ACF, 2010 (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

Appendix D.3A: Description of supplemental findings for the social-emotional development domain, follow-up findings

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (4-year-old cohort)								
Social Competencies Checklist	4-year-old cohort, Kindergarten	1,556	11.10 (1.33)	11.17 (1.05)	–0.07	–0.06	–2	.38
Social Skills and Positive Approaches to Learning Scale	4-year-old cohort, Kindergarten	1,556	12.66 (1.63)	12.63 (1.53)	0.03	0.02	+1	.78
Total Problem Behavior Scale (Negative)	4-year-old cohort, Kindergarten	1,555	5.18 (3.88)	4.99 (3.29)	–0.19	–0.05	–2	.46
Adjustment Scales for Preschool Intervention (ASPI), Aggressive Behavior Dimension (Negative)	4-year-old cohort, Grade 1	1,166	48.56 (7.42)	49.12 (7.88)	0.56	0.07	+3	.38
ASPI, Inattentive/ Hyperactive Dimension (Negative)	4-year-old cohort, Grade 1	1,179	50.35 (8.47)	50.50 (8.22)	0.15	0.02	+1	.85
ASPI, Oppositional Dimension (Negative)	4-year-old cohort, Grade 1	1,176	47.79 (7.49)	47.88 (7.33)	0.09	0.01	0	.91
ASPI, Problems with Peer Interaction Dimension (Negative)	4-year-old cohort, Grade 1	1,226	51.33 (10.99)	51.53 (11.42)	0.20	0.02	+1	.80
ASPI, Problems with Structured Learning Dimension (Negative)	4-year-old cohort, Grade 1	1,226	51.03 (10.78)	50.29 (10.68)	–0.74	–0.07	–3	.31
ASPI, Problems with Teacher Interaction Dimension (Negative)	4-year-old cohort, Grade 1	1,226	50.14 (10.38)	48.81 (10.14)	–1.33	–0.13	–5	.11
ASPI, Shy/Socially Reticent Dimension (Negative)	4-year-old cohort, Grade 1	1,182	48.00 (7.74)	46.76 (7.35)	–1.24	–0.16	–6	.04
ASPI, Withdrawn/ Low Energy Behavior Dimension (Negative)	4-year-old cohort, Grade 1	1,177	49.87 (7.59)	49.12 (7.88)	–0.65	–0.09	–4	.26
Social Competencies Checklist	4-year-old cohort, Grade 1	1,576	11.09 (1.39)	11.13 (1.17)	–0.04	–0.03	–1	.53
Social Skills and Positive Approaches to Learning Scale	4-year-old cohort, Grade 1	1,577	12.64 (1.66)	12.63 (1.57)	0.01	0.01	0	.93
TPBS (Negative)	4-year-old cohort, Grade 1	1,577	4.84 (3.83)	5.05 (3.79)	0.21	0.06	+2	.45

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Puma et al. (2012)[a] (4-year-old cohort)								
Social Competencies Checklist	4-year-old cohort, Grade 3	1,156	0.02 (1.02)	0.12 (1.00)	–0.10	–0.10	–4	.19
Social Skills and Positive Approaches to Learning Scale	4-year-old cohort, Grade 3	1,508	11.95 (1.98)	12.11 (1.91)	–0.16	–0.08	–3	.21
TPBS (Negative)	4-year-old cohort, Grade 3	1,508	5.70 (4.15)	6.18 (4.19)	0.48	0.12	+5	.14

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) and Puma et al. (2012) (4-year-old cohort), the *p*-values presented here were reported in the original study. A correction for multiple comparisons was needed and resulted in a WWC-computed critical *p*-value of .003 for ASPI, Shy/Socially Reticent Dimension (Negative) in grade 1; therefore, the WWC does not find the result to be statistically significant. For Puma et al. (2012), sample sizes and standard deviations were obtained in an author query.

Appendix D.3B: Description of supplemental findings for the social-emotional development domain, based on mothers' race/ethnicity

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (3-year-old cohort)								
Social Competencies Checklist	3-year-old cohort, White/Other, Head Start year	667	11.02 (1.40)	10.80 (1.24)	0.22	0.16	+7	.07
Social Skills and Positive Approaches to Learning Scale	3-year-old cohort, White/Other, Head Start year	667	12.55 (1.75)	12.20 (1.70)	0.35	0.20	+8	.02
Total Problem Behavior Scale (TPBS; Negative)	3-year-old cohort, White/Other, Head Start year	667	5.50 (3.56)	6.58 (3.74)	1.08	0.30	+12	< .01
Social Competencies Checklist	3-year-old cohort, Black, Head Start year	731	10.72 (1.40)	11.03 (1.24)	–0.31	–0.23	–9	.01
Social Skills and Positive Approaches to Learning Scale	3-year-old cohort, Black, Head Start year	731	12.31 (1.75)	12.34 (1.70)	–0.03	–0.02	–1	.83
TPBS (Negative)	3-year-old cohort, Black, Head Start year	731	5.48 (3.56)	5.52 (3.74)	0.04	0.01	0	.92
Social Competencies Checklist	3-year-old cohort, Hispanic, Head Start year	663	11.13 (1.40)	11.18 (1.24)	–0.05	–0.04	–1	.66
Social Skills and Positive Approaches to Learning Scale	3-year-old cohort, Hispanic, Head Start year	664	12.36 (1.75)	12.62 (1.70)	–0.26	–0.15	–6	.18
TPBS (Negative)	3-year-old cohort, Hispanic, Head Start year	664	6.45 (3.56)	6.61 (3.74)	0.16	0.04	+2	.65
DHHS ACF (2010)[b] (4-year-old cohort)								
Social Competencies Checklist	4-year-old cohort, White/Other, Head Start year	596	11.10 (1.46)	11.01 (1.19)	0.09	0.07	+3	.52
Social Skills and Positive Approaches to Learning Scale	4-year-old cohort, White/Other, Head Start year	594	12.64 (1.79)	12.53 (1.64)	0.11	0.06	+3	.58
TPBS (Negative)	4-year-old cohort, White/Other, Head Start year	595	5.57 (3.83)	5.87 (3.33)	0.30	0.08	+3	.51
Adjustment Scales for Preschool Intervention (ASPI), Aggressive Behavior Dimension (Negative)	4-year-old cohort, White/Other, Kindergarten	414	49.08 (7.22)	48.52 (7.52)	–0.56	–0.08	-3	.53

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
ASPI, Inattentive/ Hyperactive Dimension (Negative)	4-year-old cohort, White/Other, Kindergarten	418	51.03 (8.28)	51.46 (8.37)	0.43	0.05	+2	.55
ASPI, Oppositional Dimension (Negative)	4-year-old cohort, White/Other, Kindergarten	417	49.03 (7.72)	47.82 (6.95)	–1.21	–0.16	–6	.17
ASPI, Problems with Peer Interaction Dimension (Negative)	4-year-old cohort, White/Other, Kindergarten	430	51.62 (10.75)	50.08 (10.70)	–1.54	–0.14	–6	.25
ASPI, Problems with Structured Learning Dimension (Negative)	4-year-old cohort, White/Other, Kindergarten	430	51.84 (10.44)	50.76 (9.79)	–1.08	–0.11	–4	.38
ASPI, Problems with Teacher Interaction Dimension (Negative)	4-year-old cohort, White/Other, Kindergarten	430	50.35 (9.59)	48.76 (9.27)	–1.59	–0.17	–7	.20
ASPI, Shy/Socially Reticent Dimension (Negative)	4-year-old cohort, White/Other, Kindergarten	420	48.04 (7.57)	45.87 (7.64)	–2.17	–0.29	–11	.05
ASPI, Withdrawn/ Low Energy Behavior Dimension (Negative)	4-year-old cohort, White/Other, Kindergarten	419	48.95 (7.23)	48.00 (6.88)	–0.95	–0.13	–5	.40
Social Competencies Checklist	4-year-old cohort, White/Other, Kindergarten	568	11.09 (1.33)	11.19 (1.05)	–0.10	–0.08	–3	.55
Social Skills and Positive Approaches to Learning Scale	4-year-old cohort, White/Other, Kindergarten	568	12.64 (1.63)	12.53 (1.53)	0.11	0.07	+3	.60
TPBS (Negative)	4-year-old cohort, White/Other, Kindergarten	567	5.57 (3.88)	5.44 (3.29)	–0.13	–0.04	–1	.67
ASPI, Aggressive Behavior Dimension (Negative)	4-year-old cohort, White/Other, Grade 1	449	48.24 (7.42)	49.00 (7.88)	0.76	0.10	+4	.36
ASPI, Inattentive/ Hyperactive Dimension (Negative)	4-year-old cohort, White/Other, Grade 1	451	50.66 (8.47)	51.13 (8.22)	0.47	0.06	+2	.63
ASPI, Oppositional Dimension (Negative)	4-year-old cohort, White/Other, Grade 1	448	47.47 (7.49)	47.46 (7.33)	–0.01	0.00	0	.99
ASPI, Problems with Peer Interaction Dimension (Negative)	4-year-old cohort, White/Other, Grade 1	464	51.04 (10.99)	51.15 (11.42)	0.11	0.01	0	.90
ASPI, Problems with Structured Learning Dimension (Negative)	4-year-old cohort, White/Other, Grade 1	464	51.96 (10.78)	50.77 (10.68)	–1.19	–0.11	–4	.33

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
ASPI, Problems with Teacher Interaction Dimension (Negative)	4-year-old cohort, White/Other, Grade 1	464	50.59 (10.38)	47.71 (10.14)	–2.88	–0.28	–11	.03
ASPI, Shy/Socially Reticent Dimension (Negative)	4-year-old cohort, White/Other, Grade 1	451	48.66 (7.74)	46.17 (7.35)	–2.49	–0.33	–13	.03
ASPI, Withdrawn/ Low Energy Behavior Dimension (Negative)	4-year-old cohort, White/Other, Grade 1	452	50.38 (7.59)	48.31 (6.95)	–2.07	–0.28	–11	.02
Social Competencies Checklist	4-year-old cohort, White/Other, Grade 1	570	11.05 (1.39)	11.12 (1.17)	–0.07	–0.05	–2	.60
Social Skills and Positive Approaches to Learning Scale	4-year-old cohort, White/Other, Grade 1	571	12.66 (1.66)	12.41 (1.57)	0.25	0.15	+6	.09
TPBS (Negative)	4-year-old cohort, White/Other, Grade 1	571	5.38 (3.83)	5.13 (3.79)	–0.25	–0.07	–3	.54
ASPI, Oppositional Dimension (Negative)	4-year-old cohort, Black, Grade 1	238	49.34 (7.49)	49.38 (7.33)	0.04	0.01	0	.98
Social Competencies Checklist	4-year-old cohort, Hispanic, Head Start	673	11.02 (1.46)	11.09 (1.19)	–0.07	–0.05	–2	.98
Social Skills and Positive Approaches to Learning	4-year-old cohort, Hispanic, Head Start	673	12.24 (1.79)	12.53 (1.64)	–0.29	–0.17	–7	.33
TPBS (Negative)	4-year-old cohort, Hispanic, Head Start	673	6.30 (3.83)	6.14 (3.33)	–0.16	–0.04	–2	.68
ASPI, Aggressive Behavior Dimension (Negative)	4-year-old cohort, Hispanic, Grade 1	485	48.10 (7.42)	48.28 (7.88)	0.18	0.02	+1	.85
ASPI, Inattentive/ Hyperactive Dimension (Negative)	4-year-old cohort, Hispanic, Grade 1	494	49.34 (8.47)	49.64 (8.22)	0.30	0.04	+1	.85
ASPI, Oppositional Dimension (Negative)	4-year-old cohort, Hispanic, Grade 1	490	46.97 (7.49)	47.42 (7.33)	0.45	0.06	+2	.76
ASPI, Shy/Socially Reticent Dimension (Negative)	4-year-old cohort, Hispanic, Grade 1	494	47.68 (7.74)	47.10 (7.35)	–0.58	–0.08	–3	.42
ASPI, Withdrawn/ Low Energy Behavior Dimension (Negative)	4-year-old cohort, Hispanic, Grade 1	491	50.06 (7.59)	49.35 (6.95)	–0.71	–0.10	–4	.33

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Social Competencies Checklist	4-year-old cohort, Hispanic, Grade 1	650	11.13 (1.39)	11.18 (1.17)	–0.05	–0.04	–2	.57
Social Skills and Positive Approaches to Learning Scale	4-year-old cohort, Hispanic, Grade 1	650	12.68 (1.66)	13.06 (1.57)	–0.38	–0.23	–9	< .01
TPBS (Negative)	4-year-old cohort, Hispanic, Grade 1	650	5.25 (3.83)	5.27 (3.79)	0.02	0.01	0	.96

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (3-year-old cohort), the *p*-values presented here were reported in the original study. A correction for multiple comparisons was needed but did not affect whether the contrast on *TPBS (Negative)* for students with White/Other mothers in the Head Start year was found to be statistically significant. The correction for multiple comparisons resulted in WWC-computed critical *p*-values of .017 for the *Social Skills and Positive Approaches to Learning Scale* for students with White/Other mothers and .011 for the *Social Competencies Checklist* for students with Black mothers during the Head Start year (the unrounded study-reported *p*-value for the Social Competencies Checklist was .012); therefore, the WWC does not find these results to be statistically significant. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

[b] For DHHS ACF (2010) (4-year-old cohort), the *p*-values presented here were reported in the original study. The study-reported *p*-value for the *ASPI Shy/Socially Reticent Dimension (Negative)* for students with White/Other mothers in kindergarten was .051; this *p*-value is not statistically significant. A correction for multiple comparisons was needed and resulted in WWC-computed critical *p*-values of .01 for the *ASPI Problems with Teacher Interaction Dimension*, .004 for the *ASPI Shy/Socially Reticent Dimension*, and .003 for the *ASPI Withdrawn-Low Energy Behavior Dimension*, for students with White/Other mothers in Grade 1, and a WWC-computed critical *p*-value of .001 for the Social Skills and Positive Approaches to Learning Scale for students with Hispanic mothers in Grade 1 (the unrounded study-reported *p*-value for the *Social Skills and Positive Approaches to Learning Scale* was .004); therefore, the WWC does not find these results to be statistically significant. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

Appendix D.3C: Description of supplemental findings for the social-emotional development domain subscales

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (3-year-old cohort)								
Total Problem Behavior Scale (TPBS), Aggressive Behavior Subscale (Negative)	3-year-old cohort, Head Start year	2,062	2.97 (1.71)	3.05 (1.73)	0.08	0.05	+2	.42
TPBS, Hyperactive Behavior Subscale (Negative)	3-year-old cohort, Head Start year	2,062	1.71 (1.51)	2.00 (1.58)	0.29	0.19	+7	< .01
TPBS, Withdrawn Behavior Subscale (Negative)	3-year-old cohort, Head Start year	2,060	0.55 (0.90)	0.58 (1.00)	0.03	0.03	+1	.71
DHHS ACF (2010)[b] (4-year-old cohort)								
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, Head Start year	1,630	2.73 (1.77)	2.86 (1.58)	0.13	0.08	+3	.26
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, Head Start year	1,630	1.71 (1.51)	1.77 (1.44)	0.06	0.04	+2	.50
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, Kindergarten	1,556	2.41 (1.82)	2.47 (1.56)	0.06	0.03	+1	.61
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, Kindergarten	1,556	1.53 (1.52)	1.39 (1.46)	–0.14	–0.09	–4	.17
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, Grade 1	1,577	2.20 (1.82)	2.29 (1.75)	0.09	0.05	+2	.48
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, Grade 1	1,577	1.43 (1.53)	1.46 (1.54)	0.03	0.02	+1	.78
TPBS, Withdrawn Behavior Subscale (Negative)	4-year-old cohort, Grade 1	1,576	0.71 (1.01)	0.83 (1.04)	0.12	0.12	+5	.08
Puma et al. (2012)[b] (4-year-old cohort)								
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, Grade 3	1,508	2.24 (1.79)	2.47 (1.81)	0.23	0.13	+5	.07
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, Grade 3	1,508	1.91 (1.65)	1.99 (1.65)	0.08	0.05	+2	.52

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (3-year-old cohort), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study.

[b] For DHHS ACF (2010) and Puma et al. (2012) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. For Puma et al. (2012), sample sizes and standard deviations were obtained in an author query.

Appendix D.3D: Description of supplemental findings for the social-emotional development domain subscales, based on mothers' race/ethnicity

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (3-year-old cohort)								
Total Problem Behavior Scale (TPBS), Aggressive Behavior Subscale (Negative)	3-year-old cohort, White/Other, Head Start year	667	2.90 (1.71)	3.19 (1.73)	0.29	0.17	+7	0.06
TPBS, Hyperactive Behavior Subscale (Negative)	3-year-old cohort, White/Other, Head Start year	667	1.58 (1.51)	2.02 (1.58)	0.44	0.29	+11	< .01
TPBS, Withdrawn Behavior Subscale (Negative)	3-year-old cohort, White/Other, Head Start year	665	0.56 (0.90)	0.69 (1.00)	0.13	0.14	+5	0.19
TPBS, Aggressive Behavior Subscale (Negative)	3-year-old cohort, Black, Head Start year	731	2.85 (1.71)	2.82 (1.73)	–0.03	–0.02	–1	.85
TPBS, Hyperactive Behavior Subscale (Negative)	3-year-old cohort, Black, Head Start year	731	1.59 (1.51)	1.69 (1.58)	0.10	0.07	+3	.52
TPBS, Withdrawn Behavior Subscale (Negative)	3-year-old cohort, Black, Head Start year	731	0.44 (0.90)	0.44 (1.00)	0.00	0.00	0	.99
TPBS, Aggressive Behavior Subscale (Negative)	3-year-old cohort, Hispanic, Head Start year	664	3.17 (1.71)	3.13 (1.73)	–0.04	–0.02	–1	.84
TPBS, Hyperactive Behavior Subscale (Negative)	3-year-old cohort, Hispanic, Head Start year	664	1.98 (1.51)	2.31 (1.58)	0.33	0.21	+8	.06
TPBS, Withdrawn Behavior Subscale (Negative)	3-year-old cohort, Hispanic, Head Start year	664	0.66 (0.90)	0.58 (1.00)	–0.08	–0.09	–3	.56
DHHS ACF (2010)[b] (4-year-old cohort)								
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, White/Other, Head Start year	596	2.71 (1.77)	2.96 (1.58)	0.25	0.15	+6	.10
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, White/Other, Head Start year	595	1.61 (1.51)	1.74 (1.44)	0.13	0.09	+3	.51
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, White/Other, Kindergarten	568	2.55 (1.82)	2.56 (1.56)	0.01	0.01	0	.95
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, White/Other, Kindergarten	568	1.48 (1.52)	1.39 (1.46)	–0.09	–0.06	–2	.64

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, White/Other, Grade 1	571	2.27 (1.82)	2.23 (1.75)	–0.04	–0.02	–1	.84
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, White/Other, Grade 1	571	1.56 (1.53)	1.44 (1.54)	–0.12	–0.08	–3	.48
TPBS, Withdrawn Behavior Subscale (Negative)	4-year-old cohort, White/Other, Grade 1	571	0.97 (1.01)	0.98 (1.04)	0.01	0.01	0	.90
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, Hispanic, Head Start year	673	3.07 (1.77)	2.87 (1.58)	–0.20	–0.12	–5	.33
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, Hispanic, Head Start year	673	2.00 (1.51)	2.01 (1.44)	0.01	0.01	0	.94
TPBS, Aggressive Behavior Subscale (Negative)	4-year-old cohort, Hispanic, Grade 1	650	2.50 (1.82)	2.48 (1.75)	–0.02	–0.01	0	.95
TPBS, Hyperactive Behavior Subscale (Negative)	4-year-old cohort, Hispanic, Grade 1	650	1.65 (1.53)	1.56 (1.54)	–0.09	–0.06	–2	.48
TPBS, Withdrawn Behavior Subscale (Negative)	4-year-old cohort, Hispanic, Grade 1	650	0.64 (1.01)	0.81 (1.04)	0.17	0.17	+7	.14

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (3-year-old cohort), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

[b] For DHHS ACF (2010) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

Appendix D.4A: Description of supplemental findings for the alphabetics domain, follow-up findings

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (4-year-old cohort)								
Letter Naming Task	4-year-old cohort, Kindergarten	1,533	22.99 (6.10)	22.65 (6.59)	0.34	0.05	+2	.35
Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP), Elision Subtest	4-year-old cohort, Kindergarten	1,534	321.89 (49.63)	323.91 (47.07)	-2.02	-0.04	-2	.59
WJ-III, Basic Reading Skills Test	4-year-old cohort, Kindergarten	1,530	404.79 (31.24)	405.39 (32.50)	-0.60	-0.02	-1	.77
WJ-III, Letter-Word Identification Subtest	4-year-old cohort, Kindergarten	1,534	378.08 (31.61)	378.15 (33.53)	-0.07	0.00	0	.97
WJ-III, Spelling Subtest	4-year-old cohort, Kindergarten	1,535	413.91 (28.61)	414.12 (29.23)	-0.21	-0.01	0	.90
WJ-III, Word Attack Subtest	4-year-old cohort, Kindergarten	1,530	431.60 (34.35)	432.68 (34.52)	-1.08	-0.03	-1	.63
WJ-III, Basic Reading Skills	4-year-old cohort, Grade 1	1,523	451.04 (32.36)	449.81 (33.13)	1.23	0.04	+2	.52
WJ-III, Letter-Word Identification Subtest	4-year-old cohort, Grade 1	1,525	433.01 (36.22)	432.26 (36.54)	0.75	0.02	+1	.73
WJ-III, Spelling Subtest	4-year-old cohort, Grade 1	1,527	451.88 (25.56)	450.13 (26.38)	1.75	0.07	+3	.44
WJ-III, Word Attack Subtest	4-year-old cohort, Grade 1	1,524	469.10 (31.13)	467.41 (32.76)	1.69	0.05	+2	.34
WJ-III, Writing Sample	4-year-old cohort, Grade 1	1,526	479.87 (13.44)	479.75 (14.29)	0.12	0.01	0	.86
Puma et al. (2012)[a] (4-year-old cohort)								
WJ-III, Letter-Word Identification Subtest	4-year-old cohort, Grade 3	1,422	482.10 (29.48)	480.60 (28.72)	1.50	0.05	+2	.45

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) and Puma et al. (2012) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. For Puma et al. (2012), sample sizes and standard deviations were obtained in an author query.

Appendix D.4B: Description of supplemental findings for the alphabetics domain, based on mothers' race/ethnicity

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (3-year-old cohort)								
Letter Naming Task	3-year-old cohort, White/Other, Head Start year	661	5.01 (7.63)	3.69 (6.64)	1.32	0.18	+7	.16
Preschool Comprehensive Test of Phonological and Print Processing (CTOPPP), Elision Subtest	3-year-old cohort, White/Other, Head Start year	670	251.02 (43.62)	239.23 (52.04)	11.79	0.25	+10	.08
WJ-III, Letter-Word Identification Subtest	3-year-old cohort, White/Other, Head Start year	668	308.04 (25.64)	301.56 (23.98)	6.48	0.26	+10	.03
WJ-III, Spelling Subtest	3-year-old cohort, White/Other, Head Start year	670	345.87 (22.19)	345.33 (22.87)	0.54	0.02	+1	.84
Letter Naming Task	3-year-old cohort, Hispanic, Head Start year	644	4.53 (7.63)	2.56 (6.64)	1.97	0.27	+11	.01
WJ-III, Spelling Subtest	3-year-old cohort, Hispanic, Head Start year	649	349.98 (22.19)	342.39 (22.87)	7.59	0.34	+13	.02
DHHS ACF (2010)[b] (4-year-old cohort)								
CTOPPP, Elision Subtest	4-year-old cohort, White/Other, Head Start year	594	290.82 (48.66)	292.74 (49.91)	–1.92	–0.04	–2	.63
WJ-III, Letter-Word Identification Subtest	4-year-old cohort, White/Other, Head Start year	593	327.79 (28.54)	322.67 (26.63)	5.12	0.18	+7	.18
WJ-III, Spelling Subtest	4-year-old cohort, White/Other, Head Start year	593	371.70 (25.04)	368.76 (25.71)	2.94	0.12	+5	.23
Letter Naming Task	4-year-old cohort, White/Other, Kindergarten	560	23.11 (6.10)	22.77 (6.59)	0.34	0.05	+2	.55
CTOPPP, Elision Subtest	4-year-old cohort, White/Other, Kindergarten	560	339.58 (49.63)	340.38 (47.07)	–0.80	–0.02	–1	.89
WJ-III, Basic Reading Skills	4-year-old cohort, White/Other, Kindergarten	559	407.61 (31.24)	409.12 (32.50)	–1.51	–0.05	–2	.70
WJ-III, Letter-Word Identification Subtest	4-year-old cohort, White/Other, Kindergarten	560	379.56 (31.61)	382.15 (33.53)	–2.59	–0.08	–3	.53

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
WJ-III, Spelling Subtest	4-year-old cohort, White/Other, Kindergarten	560	412.60 (28.61)	415.94 (29.23)	–3.34	–0.12	–5	.20
WJ-III, Word Attack Subtest	4-year-old cohort, White/Other, Kindergarten	559	435.77 (34.35)	436.08 (34.52)	–0.31	–0.01	0	.94
WJ-III, Basic Reading Skills	4-year-old cohort, White/Other, Grade 1	547	454.80 (32.36)	453.78 (33.13)	1.02	0.03	+1	.78
WJ-III, Letter-Word Identification Subtest	4-year-old cohort, White/Other, Grade 1	548	437.50 (36.22)	436.98 (36.54)	0.52	0.01	+1	.90
WJ-III, Spelling Subtest	4-year-old cohort, White/Other, Grade 1	548	453.23 (25.56)	453.54 (26.38)	–0.31	–0.01	0	.93
WJ-III, Word Attack Subtest	4-year-old cohort, White/Other, Grade 1	547	472.11 (31.13)	470.62 (32.76)	1.49	0.05	+2	.66
WJ-III, Writing Sample	4-year-old cohort, White/Other, Grade 1	547	479.89 (13.44)	480.12 (14.29)	–0.23	–0.02	–1	.87
WJ-III, Basic Reading Skills Test	4-year-old cohort, Black, Kindergarten	337	407.72 (31.24)	401.13 (32.50)	6.59	0.21	+8	.06
WJ-III, Word Attack Subtest	4-year-old cohort, Black, Kindergarten	337	432.89 (34.35)	427.56 (34.52)	5.33	0.15	+6	.13
Letter Naming Task	4-year-old cohort, Hispanic, Head Start year	675	8.11 (9.83)	6.99 (9.41)	1.12	0.12	+5	.29
CTOPPP, Elision Subtest	4-year-old cohort, Hispanic, Head Start year	666	252.49 (48.66)	249.15 (49.91)	3.34	0.07	+3	.53
WJ-III, Letter-Word Identification Subtest	4-year-old cohort, Hispanic, Head Start year	677	315.30 (28.54)	311.81 (26.63)	3.49	0.13	+5	.41
WJ-III, Spelling Subtest	4-year-old cohort, Hispanic, Head Start year	680	370.73 (25.04)	368.37 (25.71)	2.36	0.09	+4	.48
WJ-III, Basic Reading Skills	4-year-old cohort, Hispanic, Grade 1	633	449.72 (32.36)	451.36 (33.13)	–1.64	–0.05	–2	.70
WJ-III, Letter-Word Identification Subtest	4-year-old cohort, Hispanic, Grade 1	634	430.46 (36.22)	433.19 (36.54)	–2.73	–0.08	–3	.57

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
WJ-III, Spelling Subtest	4-year-old cohort, Hispanic, Grade 1	634	450.26 (25.56)	449.32 (26.38)	0.94	0.04	+1	.75
WJ-III, Word Attack Subtest	4-year-old cohort, Hispanic, Grade 1	633	469.07 (31.13)	469.60 (32.76)	-0.53	-0.02	-1	.89
WJ-III, Writing Sample	4-year-old cohort, Hispanic, Grade 1	634	480.96 (13.44)	480.74 (14.29)	0.22	0.02	+1	.88

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (3-year-old cohort), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

[b] For DHHS ACF (2010) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

Appendix D.5A: Description of supplemental findings for the cognition domain, follow-up findings

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (4-year-old cohort)								
WJ-III, Pre-Academic Skills	4-year-old cohort, Kindergarten	1,533	406.23 (22.61)	406.48 (24.25)	–0.25	–0.01	0	.87
WJ-III, Academic Applications	4-year-old cohort, Grade 1	1,524	461.77 (16.92)	461.22 (16.70)	0.55	0.03	+1	.61
WJ-III, Academic Skills	4-year-old cohort, Grade 1	1,517	449.02 (23.70)	447.71 (24.70)	1.31	0.05	+2	.38
WJ-III, Pre-Academic Skills	4-year-old cohort, Grade 1	1,525	446.66 (24.32)	445.44 (24.99)	1.22	0.05	+2	.41

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study.

Appendix D.5B: Description of supplemental findings for the cognition domain, based on mothers' race/ethnicity

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (3-year-old cohort)								
WJ-III, Pre-Academic Skills	3-year-old cohort, White/Other, Head Start year	663	347.30 (19.53)	343.20 (19.72)	4.10	0.21	+8	.06
Color Naming/ Identification	3-year-old cohort, White/Other, Head Start year	670	0.61 (0.50)	0.62 (0.50)	–0.01	–0.02	–1	.98
DHHS ACF (2010)[b] (4-year-old cohort)								
WJ-III, Pre-Academic Skills	4-year-old cohort, White/Other, Head Start year	593	368.68 (21.14)	365.31 (21.90)	3.37	0.16	+6	.13
Color Naming/ Identification	4-year-old cohort, White/Other, Head Start year	593	0.81 (0.44)	0.77 (0.48)	0.04	0.09	+3	.43
WJ-III, Pre-Academic Skills	4-year-old cohort, White/Other, Kindergarten	560	407.87 (22.61)	410.16 (24.25)	–2.29	–0.10	–4	.29
WJ-III, Academic Applications	4-year-old cohort, White/Other, Grade 1	546	464.89 (16.92)	463.40 (16.70)	1.49	0.09	+4	.38
WJ-III, Academic Skills	4-year-old cohort, White/Other, Grade 1	546	450.62 (23.70)	450.42 (24.70)	0.20	0.01	0	.95
WJ-III, Pre-Academic Skills	4-year-old cohort, White/Other, Grade 1	548	450.15 (24.32)	449.47 (24.99)	0.68	0.03	+1	.81
WJ-III, Pre-Academic Skills	4-year-old cohort, Hispanic, Head Start year	672	358.58 (21.14)	355.09 (21.90)	3.49	0.16	+6	.35
Color Naming/ Identification	4-year-old cohort, Hispanic, Head Start year	679	0.62 (0.44)	0.56 (0.48)	0.06	0.13	+5	.27
WJ-III, Academic Applications	4-year-old cohort, Hispanic, Grade 1	634	460.19 (16.92)	460.71 (16.70)	–0.52	–0.03	–1	.75
WJ-III, Academic Skills	4-year-old cohort, Hispanic, Grade 1	632	447.92 (23.70)	448.47 (24.70)	–0.55	–0.02	–1	.85
WJ-III, Pre-Academic Skills	4-year-old cohort, Hispanic, Grade 1	634	444.35 (24.32)	444.83 (24.99)	–0.48	–0.02	–1	.87

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (3-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

[b] For DHHS ACF (2010) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

Appendix D.6A: Description of supplemental findings for the comprehension domain, follow-up findings

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (4-year-old cohort)								
Peabody Picture Vocabulary Test, Third Edition–Adapted (PPVT-Adapted)	4-year-old cohort, Kindergarten	1,535	334.21 (39.08)	331.85 (41.20)	2.36	0.06	+2	.40
PPVT-Adapted	4-year-old cohort, Grade 1	1,527	363.07 (32.18)	358.74 (32.18)	4.33	0.13	+5	.08
WJ-III, Passage Comprehension Subtest	4-year-old cohort, Grade 1	1,524	450.28 (24.96)	449.86 (23.85)	0.42	0.02	+1	.81
Puma et al. (2012)[a] (4-year-old cohort)								
PPVT-Adapted	4-year-old cohort, Grade 3	1,422	408.14 (29.83)	405.74 (28.65)	2.40	0.08	+3	.30

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) and Puma et al. (2012) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. For Puma et al. (2012), sample sizes and standard deviations were obtained in an author query.

Appendix D.6B: Description of supplemental findings for the comprehension domain, based on mothers' race/ethnicity

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (3-year-old cohort)								
Peabody Picture Vocabulary Test, Third Edition–Adapted (PPVT-Adapted)	3-year-old cohort, White/Other, Head Start year	670	271.51 (34.26)	265.88 (37.03)	5.63	0.16	+6	.05
DHHS ACF (2010)[b] (4-year-old cohort)								
PPVT-Adapted	4-year-old cohort, White/Other, Head Start year	594	315.47 (35.91)	314.60 (39.76)	0.87	0.02	+1	.75
PPVT-Adapted	4-year-old cohort, White/Other, Kindergarten	560	357.65 (39.08)	360.09 (41.20)	–2.44	–0.06	–2	.37
PPVT-Adapted	4-year-old cohort, White/Other, Grade 1	548	383.65 (32.18)	379.75 (32.18)	3.90	0.12	+5	.10
WJ-III, Passage Comprehension Subtest	4-year-old cohort, White/Other, Grade 1	546	454.97 (24.96)	451.93 (23.85)	3.04	0.12	+5	.28
PPVT-Adapted	4-year-old cohort, Hispanic, Head Start year	674	273.80 (35.91)	266.04 (39.76)	7.76	0.21	+8	.10
PPVT-Adapted	4-year-old cohort, Hispanic, Grade 1	634	344.96 (32.18)	342.76 (32.18)	2.20	0.07	+3	.59
WJ-III, Passage Comprehension Subtest	4-year-old cohort, Hispanic, Grade 1	634	447.20 (24.96)	449.36 (23.85)	–2.16	–0.09	–4	.34

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (3-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The unrounded study-reported *p*-value for the *PPVT-Adapted* outcome was .046, which is statistically significant. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

[b] For DHHS ACF (2010) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

Appendix D.7A: Description of supplemental findings for the language development domain, follow-up findings

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (4-year-old cohort)								
WJ-III, Oral Comprehension Subtest	4-year-old cohort, Kindergarten	1,533	456.52 (19.18)	457.29 (17.74)	–0.77	–0.04	–2	.55
WJ-III, Oral Comprehension Subtest	4-year-old cohort, Grade 1	1,526	473.42 (17.92)	472.36 (16.99)	1.06	0.06	+2	.44

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study.

Appendix D.7B: Description of supplemental findings for the language development domain, based on mothers' race/ethnicity

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
DHHS ACF (2010)[a] (3-year-old cohort)								
WJ-III, Oral Comprehension Subtest	3-year-old cohort, White/Other, Head Start year	666	443.59 (14.09)	440.46 (14.00)	3.13	0.22	+9	.01
DHHS ACF (2010)[b] (4-year-old cohort)								
WJ-III, Oral Comprehension Subtest	4-year-old cohort, White/Other, Head Start year	593	452.91 (17.77)	454.06 (18.09)	–1.15	–0.06	–3	.59
WJ-III, Oral Comprehension Subtest	4-year-old cohort, White/Other, Kindergarten	559	467.23 (19.18)	468.10 (17.74)	–0.87	–0.05	–2	.49
WJ-III, Oral Comprehension Subtest	4-year-old cohort, White/Other, Grade 1	548	482.54 (17.92)	481.50 (16.99)	1.04	0.06	+2	.46
WJ-III, Oral Comprehension Subtest	4-year-old cohort, Hispanic, Grade 1	634	464.26 (17.92)	464.38 (16.99)	–0.12	–0.01	0	.95

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding.

[a] For DHHS ACF (2010) (3-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

[b] For DHHS ACF (2010) (4-year-old cohort), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The findings in this table are from subgroup analyses based on the biological mothers/caregivers' race/ethnicity.

Endnotes

* On March 31, 2017 the WWC modified this report in response to a set of questions received by the WWC Help Desk. Based on the questions, the WWC added findings based on mothers' race/ethnicity in the mathematics achievement, social-emotional development, alphabetics, cognition, comprehension, and language development domains. These additional comparisons are presented as supplemental findings in the appendix and do not factor into the intervention's ratings of effectiveness. The WWC has not added studies to the evidence base, updated the literature search, or changed any study ratings since the July 2015 report.

[1] The descriptive information for this program was obtained from a publicly available source: the *Head Start* website (http://www.acf.hhs.gov/programs/ohs, downloaded June 2014). The WWC requests that program offices review the program description sections for accuracy from their perspective. The program description was provided to the program office in March 2014, and the WWC incorporated feedback from the office. Further verification of the accuracy of the descriptive information for this program is beyond the scope of this review.

[2] At least 90% of enrolled children must be from families with incomes below the federal poverty level, receiving public assistance, or are homeless. Children may also be enrolled if they are in foster care. In addition, *Head Start* programs must reserve 10% of enrollment slots for children with diagnosed disabilities.

[3] This intervention report focuses on the effectiveness of the three most common models of *Head Start*: center-based programs with home visits, family child care programs, and home visits, and any combination of these three models. There are other types of *Head Start* programs, including *Early Head Start*, *Migrant and Seasonal Head Start*, and *Family and Community Partnerships Head Start*; these alternatives were not reviewed for this intervention report. This intervention report does not include research assessing the effectiveness of particular curricula used in *Head Start*; rather, this intervention report focuses on the effectiveness of attending a *Head Start* program.

[4] The literature search reflects documents publicly available by July 2014. The studies in this report were reviewed using the Standards from the WWC Procedures and Standards Handbook (version 3.0), along with those described in the Early Childhood Education review protocol (version 3.0). The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available. A quick review blast of DHHS ACF (2010) was released and revised in 2010, which rated the study as *meets WWC group design standards with reservations* for both cohorts due to high attrition. Based on analytic sample sizes from the series of reports, there is an analysis which has low attrition, resulting in a current study rating of *meets WWC group design standards without reservations*. The quick review blast included ratings for follow-up comparisons; this intervention report presents follow-up findings as supplemental findings that do not factor into the intervention's rating of effectiveness. Additionally, all follow-up comparisons for the 3-year-old cohort *do not meet WWC group design standards* and are not presented in the supplemental findings. For these analyses, approximately 50% of the children originally assigned to the comparison group enrolled in *Head Start* as 4-year-olds. Finally, though the quick review blast focused only on academic outcomes, this intervention report includes outcomes in additional domains. In the 2010 quick review blast, DHHS ACF (2010) was cited as Puma et al. (2010).

[5] Absence of conflict of interest: This intervention report included studies conducted by staff from Abt Associates Inc. Because Abt Associates Inc. is one of the contractors that administers the WWC, those studies were rated by staff members from a different organization. The report was reviewed by the lead methodologist, a WWC Quality Assurance reviewer, and an external peer reviewer.

[6] DHHS ACF (2010) also presents follow-up findings from kindergarten, grade 1, and grade 3 in the general reading achievement, mathematics achievement, social-emotional development, alphabetics, cognition, comprehension, and language development domains; subscale findings from the social-emotional development domain; and findings based on mothers' race/ethnicity in the mathematics achievement, social-emotional development, alphabetics, cognition, comprehension, and language development domains. These additional comparisons are presented as supplemental findings in the appendix and do not factor into the intervention's ratings of effectiveness.

[7] For criteria used in the determination of the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 50. These improvement index numbers show the average and range of individual-level improvement indices for all findings across the studies.

[8] In this report, "grantees" refer to organizations that administered and had financial responsibility for programs, and "delegates" refer to organizations that were subcontracted to grantees to administer programs.

[9] The number of grantees and *Head Start* centers is based on both the 3-year-old and 4-year old cohorts. The total number of grantees and centers indicated in the report may not reflect the total number of grantees and centers that actually participated for either cohort in each study. The sample of Puerto Rican students was analyzed separately in each study; these analyses were not reviewed for inclusion in this intervention report, as the assessments were provided only in Spanish, which is outside the scope of review for the Early Childhood Education topic area (version 3.0).

[10] These numbers do not include children in Puerto Rico (see Endnote 9).

[11] This percentage includes *Head Start* programs in Puerto Rico.

[12] These numbers do not include children in Puerto Rico (see Endnote 9).

[13] This percentage includes *Head Start* programs in Puerto Rico.

[14] The review protocol stipulates children must be between 3 and 5 years of age at the time of the intervention and prioritizes the immediate posttest for determination of effectiveness. Outcomes collected when students are older—for example, in elementary or middle school—are eligible for review and included as supplemental findings in this intervention report. Note that additional outcomes were measured for the cohort of 3-year-old children in 2004, 2005, 2006, and 2008 to reflect additional follow-up outcomes at preschool, kindergarten, first grade, and third grade, respectively. However, approximately 50% of the 3-year-old children who were originally assigned to the comparison group eventually enrolled in *Head Start* as 4-year-olds. The WWC has determined that follow-up contrasts for the 3-year-old cohort collected after more than 1 year of *Head Start* are not included in this review, as the research design did not maintain the desired contrast after the first year.

[15] According the WWC Procedures and Standards Handbook (version 3.0), only composite test measures can contribute to the rating for the intervention when both composite test measures and their components are reported (Handbook, p. 17). Comparisons on subtest and subscale outcomes that meet WWC group design standards with or without reservations are still included in this intervention report as supplemental findings.

[16] Many of the follow-up, subscale, and race/ethnicity subgroup comparisons do not meet WWC group design standards, either because the outcomes do not meet review requirements or because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated. These comparisons are not included in Appendix D.

[17] In addition to the general reading achievement, mathematics achievement, and social-emotional development domains, the supplemental findings include comparisons in the alphabetics, cognition, comprehension, and language development domains. The authors reported, and the WWC confirmed, a statistically significant and positive effect of *Head Start* on three comparisons in the alphabetics domain and one comparison in the comprehension and language development domains, based on subgroups formed by mothers' race/ethnicity, for the 3-year-old cohort at the end of the intervention year.

[18] This is a composite measure of three subscales: (a) TPBS, Aggressive Behavior subscale, (b) TPBS, Hyperactive Behavior subscale, and (c) TPBS, Withdrawn Behavior subscale. DHHS ACF (2010) also reported comparisons on the three subscales, which are presented as supplemental findings in Appendix D.

[19] The procedure for classifying an effect based on multiple univariate outcomes within a single domain can be found in the WWC Procedures and Standards Handbook (version 3.0), Table IV.2 (p. 26).

[20] The WWC identified one additional source related to DHHS ACF (2010). The study does not contribute unique information to Appendix A.1 and is not listed here.

[21] *Head Start* programs in Puerto Rico were included in the original sample, but were analyzed separately and not included in this report. See Endnote 9 for additional details.

[22] Agencies were stratified by: geographic proximity, program percentage of Hispanic and African-American children; region; location (e.g., urban, rural, etc.); program auspice (i.e., whether programs were based in schools); whether programs were part-day only, full-day only, or both; and the percentage of a program's enrollment comprised of entering 3-year-old children.

[23] For more information on the FACES study, see http://www.acf.hhs.gov/programs/opre/research/project/head-start-family-and-child-experiences-survey-faces.

[24] This percentage includes *Head Start* programs in Puerto Rico.

[25] This percentage includes *Head Start* programs in Puerto Rico.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2015, July). *Early Childhood Education intervention report: Head Start*. Retrieved from http://whatworks.ed.gov

WWC Rating Criteria

Criteria used to determine the rating of a study

Study rating	Criteria
Meets WWC group design standards without reservations	A study that provides strong evidence for an intervention's effectiveness, such as a well-implemented RCT.
Meets WWC group design standards with reservations	A study that provides weaker evidence for an intervention's effectiveness, such as a QED or an RCT with high attrition that has established equivalence of the analytic samples.

Criteria used to determine the rating of effectiveness for an intervention

Rating of effectiveness	Criteria
Positive effects	Two or more studies show statistically significant positive effects, at least one of which met WWC group design standards for a strong design, AND No studies show statistically significant or substantively important negative effects.
Potentially positive effects	At least one study shows a statistically significant or substantively important positive effect, AND No studies show a statistically significant or substantively important negative effect AND fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects.
Mixed effects	At least one study shows a statistically significant or substantively important positive effect AND at least one study shows a statistically significant or substantively important negative effect, but no more such studies than the number showing a statistically significant or substantively important positive effect, OR At least one study shows a statistically significant or substantively important effect AND more studies show an indeterminate effect than show a statistically significant or substantively important effect.
Potentially negative effects	One study shows a statistically significant or substantively important negative effect and no studies show a statistically significant or substantively important positive effect, OR Two or more studies show statistically significant or substantively important negative effects, at least one study shows a statistically significant or substantively important positive effect, and more studies show statistically significant or substantively important negative effects than show statistically significant or substantively important positive effects.
Negative effects	Two or more studies show statistically significant negative effects, at least one of which met WWC group design standards for a strong design, AND No studies show statistically significant or substantively important positive effects.
No discernible effects	None of the studies shows a statistically significant or substantively important effect, either positive or negative.

Criteria used to determine the extent of evidence for an intervention

Extent of evidence	Criteria
Medium to large	The domain includes more than one study, AND The domain includes more than one school, AND The domain findings are based on a total sample size of at least 350 students, OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.
Small	The domain includes only one study, OR The domain includes only one school, OR The domain findings are based on a total sample size of fewer than 350 students, AND, assuming 25 students in a class, a total of fewer than 14 classrooms across studies.

Glossary of Terms

Attrition Attrition occurs when an outcome variable is not available for all participants initially assigned to the intervention and comparison groups. The WWC considers the total attrition rate and the difference in attrition rates across groups within a study.

Clustering adjustment If intervention assignment is made at a cluster level and the analysis is conducted at the student level, the WWC will adjust the statistical significance to account for this mismatch, if necessary.

Confounding factor A confounding factor is a component of a study that is completely aligned with one of the study conditions, making it impossible to separate how much of the observed effect was due to the intervention and how much was due to the factor.

Design The design of a study is the method by which intervention and comparison groups were assigned.

Domain A domain is a group of closely related outcomes.

Effect size The effect size is a measure of the magnitude of an effect. The WWC uses a standardized measure to facilitate comparisons across studies and outcomes.

Eligibility A study is eligible for review and inclusion in this report if it falls within the scope of the review protocol and uses either an experimental or matched comparison group design.

Equivalence A demonstration that the analytic sample groups are similar on observed characteristics defined in the review area protocol.

Extent of evidence An indication of how much evidence supports the findings. The criteria for the extent of evidence levels are given in the WWC Rating Criteria on p. 50.

Improvement index Along a percentile distribution of individuals, the improvement index represents the gain or loss of the average individual due to the intervention. As the average individual starts at the 50th percentile, the measure ranges from –50 to +50.

Intervention An educational program, product, practice, or policy aimed at improving student outcomes.

Intervention report A summary of the findings of the highest-quality research on a given program, product, practice, or policy in education. The WWC searches for all research studies on an intervention, reviews each against design standards, and summarizes the findings of those that meet WWC design standards.

Multiple comparison adjustment When a study includes multiple outcomes or comparison groups, the WWC will adjust the statistical significance to account for the multiple comparisons, if necessary.

Quasi-experimental design (QED) A quasi-experimental design (QED) is a research design in which study participants are assigned to intervention and comparison groups through a process that is not random.

Randomized controlled trial (RCT) A randomized controlled trial (RCT) is an experiment in which eligible study participants are randomly assigned to intervention and comparison groups.

Rating of effectiveness The WWC rates the effects of an intervention in each domain based on the quality of the research design and the magnitude, statistical significance, and consistency in findings. The criteria for the ratings of effectiveness are given in the WWC Rating Criteria on p. 50.

Single-case design A research approach in which an outcome variable is measured repeatedly within and across different conditions that are defined by the presence or absence of an intervention.

Glossary of Terms

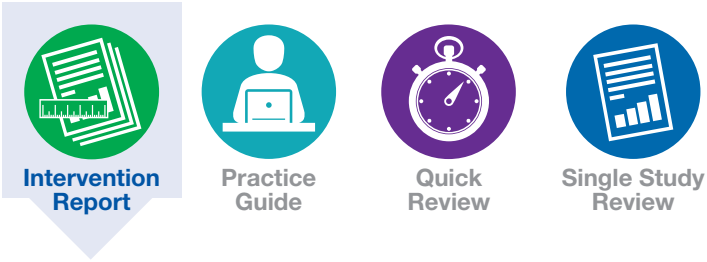
Standard deviation The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample tend to be spread out over a large range of values.

Statistical significance Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ($p < .05$).

Substantively important A substantively important finding is one that has an effect size of 0.25 or greater, regardless of statistical significance.

Systematic review A review of existing literature on a topic that is identified and reviewed using explicit methods. A WWC systematic review has five steps: 1) developing a review protocol; 2) searching the literature; 3) reviewing studies, including screening studies for eligibility, reviewing the methodological quality of each study, and reporting on high quality studies and their findings; 4) combining findings within and across studies; and, 5) summarizing the review.

Please see the WWC Procedures and Standards Handbook (version 3.0) for additional details.



An **intervention report** summarizes the findings of high-quality research on a given program, practice, or policy in education. The WWC searches for all research studies on an intervention, reviews each against evidence standards, and summarizes the findings of those that meet standards.

This intervention report was prepared for the WWC by Mathematica Policy Research under contract ED-IES-13-C-0010.