Are Ratings from Tiered Quality Rating and Improvement Systems Valid Measures of Program Quality? A Synthesis of Validation Studies from Race to the Top-Early Learning Challenge States

April 2019

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

U.S. Department of Education

Are Ratings from Tiered Quality Rating and Improvement Systems Valid Measures of Program Quality? A Synthesis of Validation Studies from Race to the Top-Early Learning Challenge States

April 2019

Lindsay Fox
Moira McCullough
Pia Caronongan
Mariesa Herrmann
Mathematica Policy Research

Tracy Rimdzius
Project Officer
Institute of Education Sciences

NCEE 2019-4001
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

U.S. Department of Education
Betsy DeVos
Secretary

Institute of Education Sciences
Mark Schneider
Director

National Center for Education Evaluation and Regional Assistance
Matthew Soldner
Commissioner

April 2019

The report was prepared for the Institute of Education Sciences under Contract No. ED-IES-10-C-0077. The project officer is Tracy Rimdzius in the National Center for Education Evaluation and Regional Assistance.

Fox, Lindsay, Moira McCullough, Pia Caronongan, and Mariesa Herrmann. (2019). *Are Ratings from Tiered Quality Rating and Improvement Systems Valid Measures of Program Quality? A Synthesis of Validation Studies from Race to the Top-Early Learning Challenge States* (NCEE 2019-4001). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

This report is available on the IES website at http://ies.ed.gov/ncee.

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

ACKNOWLEDGMENTS

Many people contributed in significant ways to this report. At Mathematica Policy Research, important contributions were made by Michael Ponza, Kimberly Boller, Lisa Dragoset, and Gretchen Kirby, who provided thoughtful, critical reviews of the report; Hanzhi Zhou, who expertly contributed to the analysis; Amanda Lee, who skillfully assisted with the interviews, Malik Mubeen, who provided excellent programming assistance; and John Kennedy, Cindy Castro, and Sharon Clark, who expertly edited and produced the report.

We would also like to extend a special thanks to the researchers who conducted the states' validation studies, who participated in interviews and responded to requests for data. Without their strong support and participation, this study would not have been possible.

CONTENTS

TABLES

FIGURES

SUMMARY

The Race to the Top—Early Learning Challenge (RTT-ELC) grants program, sponsored by the U.S. Departments of Education and Health and Human Services, aimed to improve children's access to high quality early care and education. RTT-ELC awarded more than $1 billion over three rounds of grants to help states develop and implement systems that rate early learning and development programs on quality and help them improve. These systems are known as tiered quality rating and improvement systems (TQRIS).

To strengthen the quality of early learning and development programs, TQRIS rate programs on quality standards and publicize the ratings of individual programs. States can use these ratings to identify low quality programs that need to improve, and parents can use the ratings to choose high quality programs for their children. However, the usefulness of the ratings for these purposes depends on how accurately they measure programs' quality, that is, their validity. A key objective of RTT-ELC was for states to study the validity of their ratings.

To inform states' continued development of TQRIS and future validation studies, this report synthesizes findings from validation studies conducted by nine states that received RTT-ELC grants. It also describes the challenges that researchers faced when conducting these studies. Based on studies from the nine states and interviews with the researchers who conducted them, the following key findings emerged:

- All nine states used external measures of quality to examine the validity of their ratings; eight used an independently collected measure of program quality and eight used at least one measure of children's outcomes.
- The ratings distinguished between programs with differing quality; higher-rated programs had higher scores on independent measures of quality. However, the overall level of quality for higher-rated programs could not be described as high based on these independent measures.
- The ratings were not related to differences in children's outcomes; children who attended higher-rated programs did not have better developmental outcomes than those attending lower-rated ones.
- Researchers from all nine states reported that their non-experimental designs limited the interpretation of findings. In addition, most researchers perceived recruiting child care providers for the studies and attaining sufficient representation across the rating levels as the most challenging aspects of the studies.

As TQRIS are refined further, include more programs, and become more fully implemented, it will be necessary to conduct additional studies that examine the validity of ratings.

I. INTRODUCTION

High quality early care and education yields significant benefits—especially for children from low-income and disadvantaged households (Dearing et al. 2009). Children who attend a high quality preschool for as little as a year can experience improvements in their language, literacy, and mathematics skills (Yoshikawa et al. 2013). To help increase access to high quality programs for children, particularly for children with high needs, Race to the Top—Early Learning Challenge (RTT-ELC) promoted progress on five objectives related to tiered quality rating and improvement systems (TQRIS) (Box 1). The U.S. Department of Education (ED) and U.S. Department of Health and Human Services (HHS) awarded the RTT-ELC grants to states through three rounds of competition. States received Round 1, 2, and 3 grants in 2012, 2013, and 2014, respectively.

Box 1. RTT-ELC's five TQRIS objectives

1. Developing and adopting a common, statewide TQRIS
2. Promoting participation in the TQRIS
3. Rating and monitoring early learning and development programs
4. Promoting access to high quality programs for children with high needs by:
   - Increasing the number of programs in the top levels of the TQRIS, and
   - Increasing the number and percentage of children with high needs who are enrolled in programs that are in the top levels
5. Validating the effectiveness of the TQRIS

The purpose of the grants was to strengthen the quality of early learning and development programs by supporting states as they develop and implement TQRIS. States created TQRIS to establish standards to define quality, rate programs based on those standards, and publicize the ratings of individual programs. The quality ratings are intended to help families select better programs for their children and to help states identify and support the improvement of low quality programs. The usefulness of the ratings for these purposes depends on whether they are valid measures of programs' quality. Assessing the validity of these ratings helps both to ensure the integrity of the system and inform improvements to the system (Kirby et al. 2015).

There are several potential approaches to validating a TQRIS, including examining the evidence underlying individual standards, assessing the reliability and accuracy of the information used to construct the ratings, and testing whether differences in ratings correspond with differences on other measures of quality and children's outcomes (Zellman and Fiene 2012). ED and HHS evaluated states that applied for RTT-ELC grants on the last approach — their plans to test whether differences in ratings correspond with differences on measures of program quality and children's outcomes.

Most validation studies of TQRIS conducted before RTT-ELC have found significant relationships between ratings and program quality.[1] However, validation studies have generally found weak evidence of a relationship between ratings and children's outcomes. Only two of five studies that examined individual states' TQRIS found a positive relationship between the ratings

[1] These studies include Malone et al. (2011), Elicker et al. (2011), Lahti et al. (2011), Bryant et al. (2001), and Norris and Dunn (2004). Studies of Colorado's TQRIS (Zellman et al. 2008) and Minnesota's TQRIS (Tout et al. 2011) did not consistently find a relationship between the ratings and external measures of program quality.

and children's outcomes.[2] Another study used existing data from multiple states to simulate programs' ratings under different states' TQRIS (Sabol et al. 2013). It found these simulated ratings had little association with students' math, prereading, language, and social skills.

To inform ongoing TQRIS development, the Institute of Education Sciences (IES) at ED initiated a study to learn about TQRIS in states that received RTT-ELC grants. The study focused on center-based early learning and development programs that served preschool-age children (these programs may have also served infants, toddlers, and school-age children). The study generated a series of reports and briefs. The first report examined progress on the first three TQRIS objectives (outlined in Box 1) by describing the development, structure, and characteristics of TQRIS in states that received Round 1 grants. The report found that states varied substantially in the ways they promoted participation in TQRIS, defined quality standards, verified that programs met the standards, and calculated ratings (Kirby et al. 2017). Future work from this study plans to examine states' progress on the fourth TQRIS objective, examining the (1) number and percentages of programs at top levels of the TQRIS, and (2) patterns of TQRIS ratings across states with different TQRIS characteristics and policies.

This report contributes to this larger study by focusing on the fifth RTT-ELC objective—validating the effectiveness of the TQRIS. It synthesizes findings across validation studies conducted by nine RTT-ELC states: California, Delaware, Massachusetts, Minnesota, Ohio, Oregon, Rhode Island, Washington, and Wisconsin (Box 2 includes a list of studies that we reviewed). Seven of these states (all except Oregon and Washington) received Round 1 RTT-ELC grants. Oregon and Wisconsin received Round 2 grants. We selected these nine states for the synthesis report because they were the first nine RTT-ELC states to complete a validation report and either publicly release the report or provide it to include in the synthesis.

This report also adds to the growing knowledge base about the validity of ratings from TQRIS. This knowledge base includes a recent report that synthesized validation studies conducted by ten states: Arizona, California, Delaware, Maryland, Massachusetts, Minnesota, Oregon, Rhode Island, Washington, and Wisconsin (Tout et al. 2017).[3] Nine of these states received RTT-ELC grants. In that report, the authors of the states' validation studies described findings from individual states and the patterns of findings across states, but they did not combine data or calculate averages across states. Like earlier validation studies, Tout et al. (2017) found evidence of an association between TQRIS ratings and measures of program quality. However, they concluded that evidence of a relationship between the ratings and children's outcomes was inconsistent.

This independent synthesis report builds on Tout et al. (2017) in several ways. First, this report combines data from several states to provide information about the magnitude and

[2] These studies include Zellman et al. (2008), Thornburg et al. (2009), Elicker et al. (2011), Tout et al. (2011), and Sabol and Pianta (2012). A study of Missouri's TQRIS found a positive relationship for social skills and behavior, but not vocabulary, early literacy, or math skills (Thornburg et al. 2009). A study of Virginia's TQRIS found a positive relationship for literacy skills (Sabol and Pianta 2012).

[3] Nine of these states examined relationships between TQRIS ratings and program quality, and seven examined relationships between ratings and children's outcomes.

statistical significance of the average associations between (1) TQRIS ratings and measures of program quality, and (2) TQRIS ratings and children's developmental outcomes across these states. Second, this report only analyzes associations between TQRIS ratings and children's outcomes that were based on children who had similar scores at the beginning of each validation study; this helps to maximize the likelihood that associations between TQRIS ratings and children's outcomes reflect differences in the quality of the learning environment across programs that receive different ratings (as opposed to differences between the types of children who attend programs with different ratings). Third, the two reports synthesize slightly different sets of states, although they have eight states in common. This report includes Ohio; Tout et al. (2017) include Arizona and Maryland.[4] Finally, to inform future validation studies about the challenges they may face when conducting these studies, this report describes the challenges that validation study authors reported, based on interview data we systematically collected from the authors.

Research questions

To inform states' continued development of TQRIS and future validation studies, we examined the following research questions for the nine RTT-ELC states:

- How did states validate their TQRIS?
- Do ratings of TQRIS reflect differences in the quality of programs (based on quality measures that researchers collected independently, outside of the TQRIS)?
- Do children who attend programs with higher ratings have better developmental outcomes than those who attend programs with lower ratings? Is there a relationship between programs' ratings and outcomes for all children, or specifically for low-income children?
- What were the most common challenges associated with conducting the states' validation studies?

To answer these questions, we reviewed state validation reports and interviewed the researchers who conducted the validation studies.

[4] This report examines a total of nine states, and Tout et al. (2017) examine a total of ten states. However, some states included in the two syntheses did not analyze both measures of program quality and children's outcomes. Therefore, this report analyzes the associations between TQRIS ratings and measures of program quality for eight states (one fewer than Tout et al. 2017) and the associations between ratings and children's outcomes for eight states (one more than Tout et al. 2017).

II. BACKGROUND ON STATES' TQRIS

TQRIS began in two states (Oklahoma and Colorado) in the late 1990s but expanded to 39 states by the end of 2016 (Build Initiative and Child Trends 2016). This report focuses on nine states that first implemented TQRIS from 2004 to 2013 (Table II.1). Most of these states had TQRIS that were still in flux when researchers started their validation studies; at the start of the studies, six states (California, Delaware, Minnesota, Rhode Island, Oregon, and Washington) had either not fully implemented their TQRIS or were in the process of changing them substantially.

TQRIS rate early learning and development programs against state-defined quality standards that include components such as licensing compliance, quality of the learning environment, and qualifications of the workforce. States define standards differently, based on their priorities, and can use different standards and measures to rate different types of programs. For example, states use different measures of the quality of the learning environment for programs operated out of the caregiver's home—that is, family child care programs—versus those that operate in an institutional setting such as a center—that is, center-based programs. Family child care programs are a specific type of home-based care. Home-based care can also include an individual or shared sitter or a relative.

States also set different standards for caring for children of different ages (such as infants, toddlers, and preschool-age children). For example, standards for caring for infants and toddlers require smaller group sizes than those for preschool-age children. States also may have policies—alternative pathways or automatic ratings—that allow eligible programs (such as Head Start programs) to be exempted from part or all of the TQRIS rating process because they have already demonstrated meeting a comparable set of standards (such as federal standards for receiving Head Start funds).

Based on meeting the state-defined standards, programs receive an overall rating level. The number of overall rating levels differs across states; in the nine RTT-ELC states in this report, rating levels ranged from 1 to 4 in two states and 1 to 5 in seven states. However, rating levels are not directly comparable across states, even among states with the same number of rating levels, because states define quality standards and calculate rating levels differently (Kirby et al. 2015). The commonality across states is that each rating level signifies higher quality than the rating level below it; for example, programs rated level 3 are expected to be higher quality than programs rated level 2.

Differences in TQRIS across the nine states may affect how states conducted their validation studies, how much their ratings distinguish between high and low quality programs, and whether there is a relationship between ratings and children's outcomes. The differences across the nine states include the following:

Types of programs that participate. The researchers who conducted states' validation studies had to select the types of programs to study; they could study only programs that states allowed to participate in their TQRIS. In two of the nine states, participation in TQRIS is voluntary for all programs. Other states require certain types of programs to participate (for example, programs that receive public funds to serve low-income children) but may allow other programs (such as family child care programs) to participate voluntarily.

The percentage of programs that participate in TQRIS varies across states and program types. The nine states did not consistently provide information about the rates of participation in TQRIS. Only four states—California, Delaware, Minnesota, and Ohio—provided information about TQRIS participation over the study period. In those states, participation rates for licensed centers (which ranged from about 20 to 70 percent) exceeded those for family child programs (which ranged from less than 3 percent to 25 percent) (see Appendix C for state-specific participation rates over the study period).

States use different measures of quality for family child care programs, compared with center-based programs. In addition, family child care programs typically have children with a mix of ages because one caregiver is working with a group of children, whereas center-based programs often group children by age.

Defining standards. States define the quality standards that programs must meet to earn each rating; those with standards that focus more on classroom-specific aspects of quality than on administration and management might find stronger relationships between ratings and children's outcomes than other states (Burchinal et al. 2016; Sabol et al. 2013). Standards can cover many different aspects of quality, but some relate more closely to what happens in the classroom, such as curriculum and the quality of the learning environment. All nine states included in the synthesis have standards in categories related to children's learning and development, though the terminology varies. Other standards pertain more to administration and management, which might not affect children's outcomes as directly. Four of the nine states (Delaware, Massachusetts, Ohio, and Oregon) have standards related to administration and management, assessing programs on characteristics such as their written operating policies and procedures and financial record-keeping (Build Initiative and Child Trends 2016).

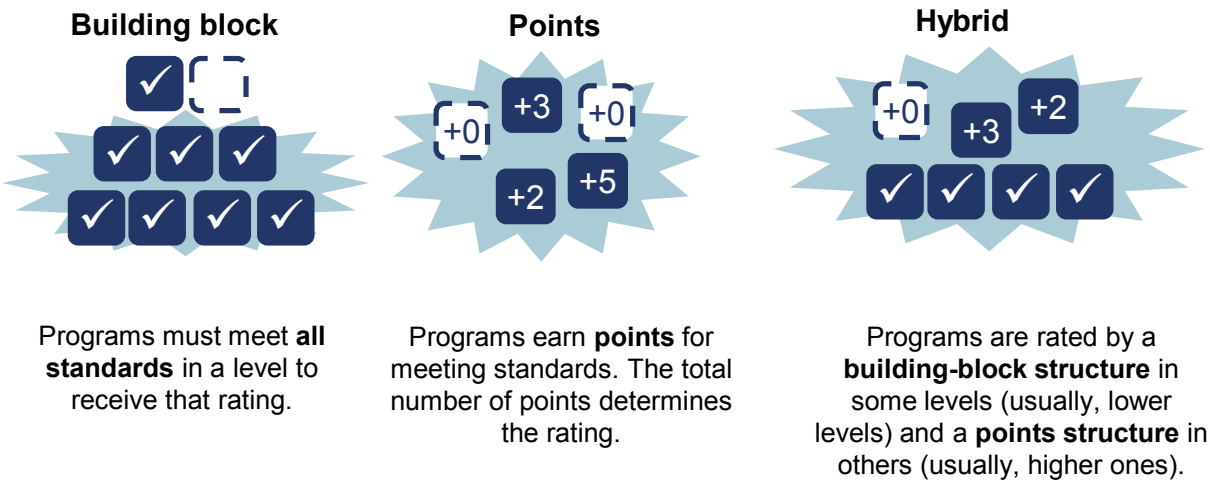
Verifying that programs meet the standards. After defining the quality standards, states must collect information to verify that programs meet these standards. States conducting validity studies had to identify additional, independent quality measures to use to examine differences across ratings. States differ in how they collect the information used to determine whether programs meet the standards. Depending on the indicator being verified, states may rely on self-reported information from programs, document reviews, and information in existing databases (Kirby et al. 2017). All of the states include an observational assessment of the classroom environment or teacher–child interactions as part of the rating process for certain rating levels.

Number of rating levels. States designing a validation study had to consider whether to examine differences across all of their individual rating levels or collapse levels into one higher-rated group and one lower-rated group. For this study, we also had to determine how to compare rating levels across states that used different numbers of levels. Based on meeting the standards, states give programs an overall rating that can range from 1 to 4 or 1 to 5. Two states have four levels, and the remaining seven have five.

Rating structure. The rating systems that determine programs' rating levels could weaken the relationship between the ratings and outcomes. TQRIS use one of three rating structures to determine a program's rating level (Figure II.1). Building block structures require programs to meet all standards within a level to receive a rating. In contrast, points and hybrid structures provide programs the flexibility to choose the standards they meet to earn a higher rating, as long

as programs receive enough points. In block systems, a program that missed qualifying for the next level based on a single standard might not be that different in quality from programs at the next level. In points and hybrid structures, programs that receive a given rating level could have met different standards. Three states use a building block structure, one state uses a points structure, and the other five states use a hybrid structure.

Figure II.1. Rating structures in TQRIS



Automatic ratings. Automatic ratings for programs that meet quality standards external to TQRIS could weaken the relationship between the ratings and outcomes. Some states automatically award high rating levels to programs that meet external quality standards (for example, programs accredited by professional organizations, such as the National Association of Education for Young Children, or Head Start programs). Automatic ratings can ease the burden of the full data collection and verification process for programs that likely would have met all the requirements for the highest rating had they gone through the full process. However, if automatically rated programs would not have obtained the highest rating through the full process, offering automatic ratings could weaken the relationship between the ratings and outcomes. Three of the nine states offer automatic ratings to accredited programs and one state offers automatic ratings to Head Start programs.

Given these differences, it is possible that the relationships between ratings and outcomes could differ across states.

Table II.1. TQRIS characteristics, by state

State	First year of operation	Types of programs that can participate	Categories on which programs are rated[a]	Number of rating levels	Rating structure	Availability of automatic rating	Use of observational measures in rating process
California[b]	2012	Voluntary	Child development and school readiness; teachers and teaching; and program and environment	5	Points	Not available	Used for rating, specific score required for points or level
Delaware	2008	Eligible programs include licensed centers	Family and community partnerships; qualifications and professional environment; management and administration; and learning environment and curriculum	5	Started as block (2008), changed to points (2012), changed to hybrid (2015)	Available for accredited programs	Used for rating, specific score required for points or level
Massachusetts	2011	Eligible programs include licensed center-based and family child care programs	Curriculum and learning; safe, healthy indoor and outdoor environments; workforce qualifications and professional development; family and community engagement; and leadership administration and management	4	Block	Not available	Used for rating, specific score required for points or level
Minnesota	2007	Voluntary	Physical health and well-being; teaching and relationships; assessment of children's progress; and teacher training and education	4	Hybrid	Available for accredited programs	Used for rating, specific score required for points or level
Ohio	2004	Eligible programs include licensed center-based and family child care programs	Learning and development; administrative and leadership practices; staff qualifications and professional development; family and community partnerships; and staff:child ratio and group size and accreditation	5	Hybrid	Not available	Used for rating, no specific score required
Oregon	2013	Eligible programs include licensed center-based and family child care programs	Children's learning and development; health and safety; personnel qualifications; family partnerships; and administrative and business practices	5	Block	Not available	Used for rating, specific score required for points or level

Table II.1. (*continued*)

State	First year of operation	Types of programs that can participate	Categories on which programs are rated[a]	Number of rating levels	Rating structure	Availability of automatic rating	Use of observational measures in rating process
Rhode Island	2009	Eligible programs include licensed center-based and family child care programs	Learning environment; minimum staff:child ratio; maximum group size; teacher qualifications; program leadership; continuous quality improvement; curriculum; child assessment; inclusive classroom practices; and family communication and involvement	5	Block	Not available	Used for rating, specific score required for points or level
Washington	2012	Eligible programs include licensed or certified center-based and family child care facilities, as well as Head Start and Washington State's pre-kindergarten programs	Children's outcomes; facility curriculum and learning environment and interactions; professional development and training, and family engagement and partnership	5	Hybrid	Available for Head Start programs	Used for rating, specific score required for points or level
Wisconsin	2010	Licensed center-based programs and family child care programs can apply	Education and training qualifications; learning environment and curriculum; professional and business practices; and children's health and well-being practices	5	Hybrid	Available for accredited programs	Used for rating, specific score required for points or level

Source: Build Initiative and Child Trends (2016).

[a] Categories are listed using state-specific terms, as recorded in the QRIS Compendium.

[b] California's TQRIS is administered locally; 16 counties in California first implemented TQRIS in 2012.

III. COMPONENTS OF VALIDATION STUDIES

To shape TQRIS that provide meaningful measures of quality, RTT-ELC states conducted studies to validate their systems. These studies could validate the TQRIS by "measuring whether the tiers of TQRIS accurately reflect different levels of program quality and whether changes in quality ratings are related to children's progress in learning, development, and kindergarten readiness" (U.S. Department of Education 2013). The process for a state's validation of the TQRIS included three main components.

Designing the study and approach to the analysis. First, researchers had to design a study that could assess whether the ratings accurately reflect differences in programs' quality and differences in children's progress in learning and development. Because children could not be randomly assigned to early childhood education programs, researchers had to use non-experimental designs to compare outcomes for children enrolled in programs with different ratings. These comparisons are meaningful only if the children are similar aside from the programs in which they enroll. Otherwise, the comparisons will simply reflect differences in family backgrounds or existing skills of children who attend programs with different ratings. To ensure that the children being compared are as similar as possible, non-experimental designs might use matching or statistical controls to adjust for differences between children. However, it is still possible that these techniques might miss an important difference between children in different programs (such as parent involvement) that is not measured but affects children's outcomes.

As part of the study design, researchers also had to decide whether to group programs into higher- and lower-rated categories for comparison or to examine differences across each individual rating level. Researchers might decide to group programs into higher- and lower-rating levels if there were not enough programs at individual rating levels to examine the ratings separately, and they might combine rating levels that were similar in quality (such as combining levels 1 and 2 and combining levels 3 and 4 in a state with four levels). The decision concerning whether to group programs into rating level categories might affect the number of programs they chose to recruit in each rating level; researchers could also make this decision during the analysis phase, after they knew how many programs they had recruited.

Recruiting programs and families to obtain a sample of sufficient size and representativeness. After finalizing the study design, researchers had to recruit programs and families to participate in the validation study. To detect differences in program quality across rating levels, researchers had to secure the participation of a sufficient number of programs that represented the various rating levels. Researchers also had to determine whether to recruit all program types, including family child care programs, or focus on center-based programs because of their prevalence. To examine differences in children's outcomes, researches also had to recruit the children enrolled at the study programs and obtain their parents' consent to administer the assessments.

Selecting measures, collecting data, and conducting the analysis. In the final stage, researchers had to select external measures of program quality and child development. They also had to administer the measures and analyze differences in outcomes on the measures across rating levels.

All components had to be completed within time frames allotted for the study. As explained earlier, at the time of the validation studies, some states had not fully implemented their TQRIS. This timing could affect researchers' schedules for recruitment and collecting data, as well as interpreting findings.

IV. DATA AND METHODS

Boxes 2 and 3 discuss the data and methods used in this report. We describe in Box 2 the data we collected from the TQRIS validation studies and from our interviews with the researchers who conducted them. We explain in Box 3 our methods for synthesizing findings across the validation studies.[5]

Box 2. Data from state validation studies and interviews with researchers

This report uses two main sources of data:

State validation studies

This report reviewed the following TQRIS validation studies from nine RTT-ELC states:

- California: Quick et al. (2016a, 2016b)
- Delaware: Karoly et al. (2016)
- Massachusetts: Roberts et al. (2016)
- Minnesota: Tout et al. (2016)
- Ohio: Heinemeier et al. (2017)
- Oregon: Lipscomb et al. (2016)
- Rhode Island: Maxwell et al. (2016)
- Washington: Soderberg et al. (2016)
- Wisconsin: Magnuson and YingChun (2015, 2016)

As we reviewed the validation reports, we documented information about the validation studies (such as types and numbers of programs included and methods used). We also recorded the statistics needed to calculate differences in program quality and child development between two groups of rating levels: high and low. For each group, these statistics include sample sizes, means and standard deviations on baseline and follow-up assessments, and regression-adjusted means or regression coefficients. If the reports did not include all of this information, we contacted the researchers to request it. We never asked researchers to conduct additional analyses.

Interviews with authors

To understand the challenges experienced by researchers conducting the TQRIS validation studies, we conducted a 30-minute semistructured phone interview with principal investigators for each of the nine states.

We first examined validation study reports and existing documentation of TQRIS validation study experiences (Lahti et al. 2013) to develop a list of specific challenges principal investigators encountered and grouped them into four broad categories:

1. Study design and analysis
2. Sample size and representativeness
3. Selecting program and child measures and collecting data
4. Study schedule and timing relative to implementing TQRIS

We then developed a semistructured interview protocol design to collect standard information across the nine studies—whether a specific challenge had been experienced and, if so, whether it was considered major or minor—along with more detailed descriptions of all challenges experienced while conducting TQRIS validation studies.

[5] Our analysis approach follows the What Works Clearinghouse (WWC) standards version 3.0 (U.S. Department of Education 2014). The WWC released version 4.0 of their standards in October 2017 after we had collected data from author queries and conducted our analyses. See Appendix B for details on the differences between the two versions as they relate to our analysis.

Box 3. Methods for synthesizing results of validation studies

This box summarizes how we synthesized results of validation studies. For more details about these methods, see Appendix B.

The nine states had different numbers of rating levels and used different methods and outcome measures in their validation studies. To combine findings across these states, we did the following:

1. **Defined two groups of rating levels: high and low.** To compare findings across states that had different numbers of rating levels (either four or five), and across states that had already combined rating levels into two groups and those that had not, we defined two groups of consistent rating levels. These groups followed the definitions used by the states that already combined rating levels:
   - High (a rating level of 3 or higher)
   - Low (a rating level of 1 or 2)
2. **Classified outcome measures into domains.** To compare findings across states that used different assessments, we classified outcome measures into domains based on the constructs they assessed. For example, the language development domain included children's scores on Letter-Word Identification on the Woodcock Johnson Tests of Achievement and the Test of Preschool Early Literacy.
3. **For each outcome measure, calculated the average difference between programs with high ratings and those with low ratings in each state.** We used different methods to calculate the standardized average differences for the program quality and child outcome measures. States also used different methods for each type of measure. For child outcome measures, measuring them at the beginning of the year (baseline) enabled researchers to try to account for differences between the types of children who attend programs with different ratings.
   a. **For program quality measures, we calculated the standardized difference.** For program quality measures, we took the average difference in program quality between higher- and lower-rated programs, and divided it by the standard deviation of this measure (calculated across all programs).
   b. **For child outcome measures, used estimates based only on similar children.** For each child development measure, we examined whether children had similar scores on the baseline assessments. We did not use outcomes if children had baseline differences that exceeded the What Works Clearinghouse (WWC) baseline equivalence standards. Among estimates that met these WWC thresholds, we used regression-adjusted estimates to calculate the standardized average difference. If those were not available, we subtracted any differences between higher- and lower-rated programs at the beginning of the year from the differences at the end of the year. In both cases, we standardized the average difference by dividing by the standard deviation.
4. **Characterized the statistical significance, sign, and magnitudes of findings.** We reported:
   - For each state, the difference between higher- and lower-rated programs, averaged across all measures in the domain, and its associated 95 percent confidence interval. This average summarizes findings within states. The confidence interval indicates a significant association if its bounds do not include zero. We followed WWC procedures to calculate these averages and confidence intervals.
   - Across states, the average difference between higher- and lower-rated programs and its associated 95 confidence interval. This average summarizes findings across states. Because the quality of programs in each rating level could differ dramatically across states, we calculated this average using a statistical approach that weighted each state by a measure of the confidence in the average (a fixed-effect meta-analysis that weights states by their inverse variance).
5. **Conducted a supplemental analysis that compared only the highest and lowest rating levels possible.** As a supplemental analysis, for states that reported disaggregated statistics by rating level, we repeated steps 1 through 4 to compare only the highest and lowest rating levels. This analysis provides information about differences in outcomes for the largest contrast of rating levels.

V. APPROACHES FOR STATES' VALIDATION STUDIES

All nine states included in this synthesis used a non-experimental design for their validation studies. However, their studies included different types of programs and used different measures and methods (Table V.1). We classified the measures into domains based on the constructs they assessed. These domains include program quality or separate types of child outcomes, such as alphabetics or social-emotional development. (See Appendix A, Table A.1 for more details on the definitions of the domains and specific measures categorized in each domain and Appendix A, Table A.2 for the specific outcome measures used for the TQRIS validation study in each state.) See Appendix C for a detailed description of each state's validation study and its findings. The descriptions in Appendix C include information that was reported systematically in all studies such as the number and types of programs in the sample (and, if available, how the sample of programs in the validation studies compare with the full sample of programs that participate in TQRIS or the full sample of programs in the state), ages of children in those programs, characteristics of the sample, and the statistical approach the authors took. For information not included in Appendix C, please refer to the individual state validation studies listed in Box 2 and cited in the "References" section.

Most states included both center-based and family child care programs, but the number of programs included varied across states. Six states included both center-based and family child care programs. Only three states (California, Massachusetts, and Rhode Island) focused entirely on center-based programs. The number of programs included in the studies ranged from about 70 (Ohio and Rhode Island) to about 300 (Minnesota and Oregon).

Three states combined center-based and family child care programs in their analyses. States use different measures of program quality for center-based and family child care programs (for example, they might use the Early Childhood Environmental Rating Scale-Revised for center-based programs and the Family Child Care Environment Rating Scale for family child care programs). However, the states that combined center-based and family child programs in their analyses argued that it was appropriate to group programs together when assessing validity of the TQRIS ratings.

Eight of the nine states used an independent measure of program quality to examine differences between programs, but they used different methods to compare programs. Only Washington did not use an independent measure of program quality to examine differences across rating levels, opting to focus solely on differences in children's outcomes across rating levels.

Researchers most commonly observed the learning environment in a subset of classrooms within programs once during the study period.[6] Drawing upon widely used rubrics, such as the Classroom Assessment Scoring System (CLASS) and the Environmental Rating Scales (ERS), they rated interactions in the learning environment such as those between staff and children; among staff, parents, and other adults; between children; and between children and the materials and activities in the classroom. (See Appendix A, Table A.3 for more information about the

[6] For some particular programs, all classrooms in the program were observed because, for example, the study observed up to four classrooms per program and the program had four or fewer classrooms.

scores and predictive validity of the rubrics.) For some rubrics, researchers also rated physical components of the learning environment, including the space, schedule, and materials that support interactions in the classroom.

Most states used one to three measures of program quality, but Ohio used seven. In some cases, the external observations were coded using the same rubrics (such as CLASS or ERS) as the observations collected for the TQRIS ratings. Four states' validation studies used the same rubric for at least one external measure, and two states' studies used a different rubric. The other two states' studies used the same rubric as the TQRIS, but argued that, in practice, none or very few of the programs in their studies had actually been rated on that rubric by the state (for example, because the state did not require observations for programs rated below the highest level). We might expect to find stronger relationships between ratings and measures of program quality that states use in their TQRIS, compared with those they do not.

Most states compared program quality outcomes for individual rating levels and did not adjust for other program characteristics. Five of the eight states that used an independent measure of quality compared programs in each rating level to programs in every other level.[7] The other three states combined programs in individual rating levels into two groups with higher and lower ratings for comparison. Two states controlled for other program characteristics in their comparisons of program quality by rating level.[8] The remaining six states simply compared average scores without conducting any statistical adjustments.

Eight states used at least one assessment of children's outcomes, but their methods for comparing programs differed. Only Oregon had not reported findings for children's outcomes at the time that we conducted this synthesis. In the other eight states, researchers collected 2 to 12 measures of children's outcomes; for each, they administered two assessments to children (once at baseline and again at a follow-up period).[9] The time between the baseline and follow-up period was less than 12 months for all states. Researchers most frequently assessed children's alphabetics, cognition, and mathematics skills, but a few states collected data on children's socioemotional development and motor skills. The assessments included the Woodcock Johnson III Letter-Word identification subscale, the Woodcock-Johnson III applied problems subscale, and the Preschool Learning Behaviors Scale.

Most states compared children's outcomes for individual rating levels and adjusted for children's performance at baseline. Only two states combined programs in individual rating levels into two groups for comparison. Six of the states accounted for children's performance at the beginning of the study when comparing their outcomes at the end of the study across rating

[7] Delaware combined the Starting with Stars and 2-star rating levels but otherwise compared individual rating levels.

[8] For comparability across states, we used findings without controls in our analyses, although it is possible that including control variables could reduce the size of the differences between higher- and lower-rated programs. See the note in Figure VI.1 for details on the two states that reported findings with controls.

[9] One state attempted to collect child development outcomes from all 3- and 4-year-old children in each site. Among the remaining states collecting child development outcomes, four states collected outcomes from a sample of children within programs, and four states collected outcomes from a sample of children within a single selected classroom or a subset of classrooms within each program.

levels. The specific child and family characteristics that researchers included as statistical controls in analyses varied widely across states; no two validation studies included the same set of covariates in their analyses. Many states (six of eight that analyzed children's outcomes) used a statistical approach (multilevel modeling) to account for the shared experiences of children within programs.

To examine whether attending a higher-rated program was beneficial for children from low-income households, four states separately analyzed outcomes of low-income children. High quality programs might be especially important for these children with fewer resources and opportunities to promote their development.

Table V.1. Summary of approaches states used to examine measures of program quality and children's outcomes

State	Number and type of programs	Program quality			Child outcomes					
		Number of external measures of program quality	Analysis combined rating levels into two groups	Controls for program characteristics	Number of child outcome measures	Analysis combined rating levels into two groups	Controls for baseline performance	Controls for other child and family characteristics	Time period for measuring change in child development	Supplemental subgroup analysis for low-income children
California	166 center-based programs	1	X		4		X	X	Fall to spring	
Delaware	156 center-based and family child care programs	3		X	5		X	X	Fall to spring	X
Massachusetts	120 center-based programs	3	X		5		X	X	Fall to spring	
Minnesota	294 center-based and family child care programs	3			6	X		X	Fall to spring	X
Ohio	72 center-based and family child care programs	7			2				Spring to fall	
Oregon	304 center-based and family child care programs	1	X		0	n.a.	n.a.	n.a.	n.a.	n.a.
Rhode Island	71 center-based programs	1			5		X	X	Fall to spring	X
Washington	100 center-based and family child care programs	0	n.a.	n.a.	12		X	X	Fall to spring	
Wisconsin	239 center-based and family child care programs	2		X	7	X	X	X	Fall to spring	X

Sources: State validation reports for California, Delaware, Massachusetts, Minnesota, Ohio, Oregon, Rhode Island, Washington, and Wisconsin.

n.a. = not applicable.

VI. ASSOCIATIONS BETWEEN RATINGS AND INDEPENDENT MEASURES OF PROGRAM QUALITY

If the ratings accurately reflect different levels of program quality, higher-rated programs should score better than lower-rated ones on quality measures collected outside of the system. Researchers from eight states tested the validity of the ratings by examining their relationship with independent assessments of the classroom environment and teacher–child interactions, such as the ERS and CLASS. Our analysis compared the outcomes of higher-rated programs (programs with rating levels of 3 or above) with those lower-rated programs (programs in the bottom two rating levels).

Higher-rated programs scored higher on independent assessments of program quality than lower-rated programs. Across all states, higher-rated programs scored 0.57 standard deviations higher on measures of program quality, on average, than lower-rated programs (shown by the blue dot in Figure VI.1). This finding across states was statistically significant (shown by the solid 95 percent confidence interval around the blue dot in Figure VI.1 that does not include zero).

In seven of eight states, higher-rated programs had significantly better scores on measures of program quality than lower-rated programs (shown by the solid 95 percent confidence intervals around the black dots in Figure VI.1 that do not include zero). In the other state (Minnesota), the average difference between higher- and lower-rated programs was positive but not significant (shown by the dashed 95 percent confidence interval around the black dot in Figure VI.1 that includes zero). Across all of the program quality measures used by each state, the range of scores was about 1.00 standard deviation; higher-rated programs scored an average of 0.20 standard deviations higher than lower-rated programs in Minnesota, and an average of 1.19 standard deviations higher in California (as shown by the black dots in Figure VI.1). The average difference in program quality varied significantly across states (based on a statistical test for heterogeneity called the "Q test").

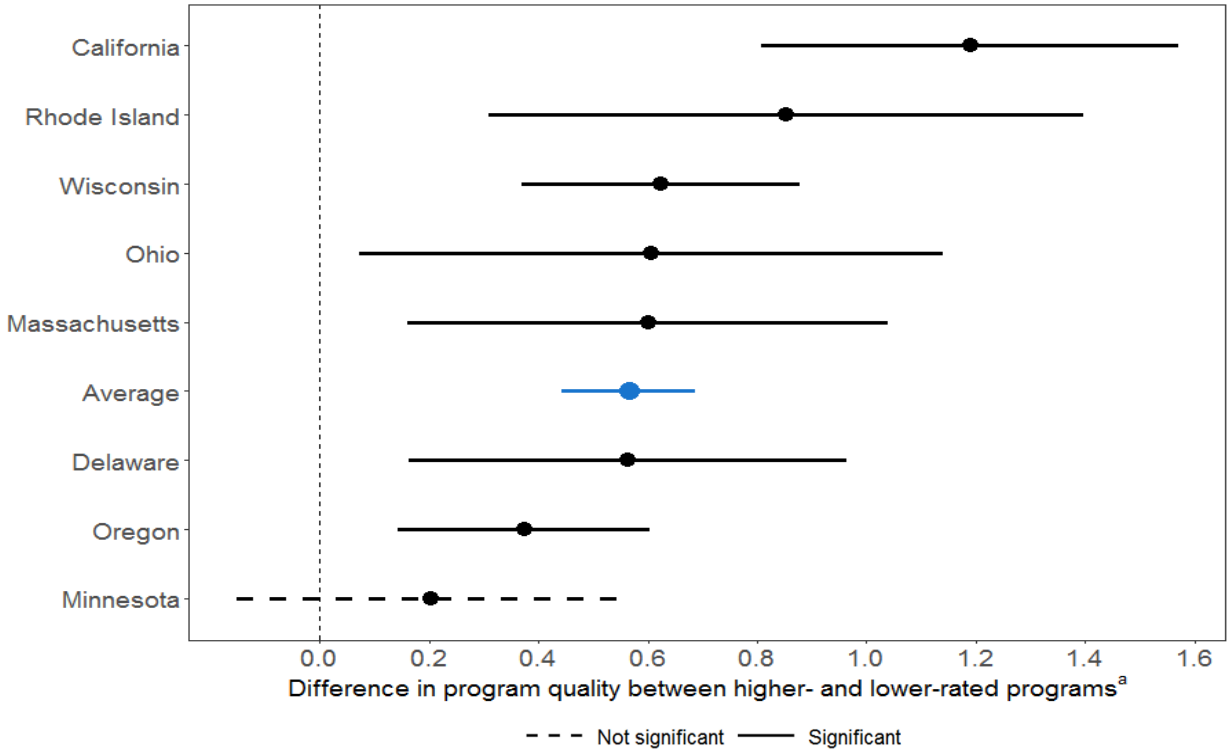
The differences in program quality scores between higher- and lower-rated programs might not be large enough to significantly increase children's learning. Based on the standard deviations of the measures that states reported, the average difference in program quality across states (0.57 standard deviations) corresponds to a difference of only about half a point on the ERS or CLASS. The largest difference (1.19 standard deviations in California) corresponds to a difference of about one point on these measures—small enough to leave higher- and lower-rated programs in the same classification of scores.

This program quality difference might not be large enough to significantly increase children's learning. For example, two recent meta-analyses found few associations between these measures and children's outcomes (Perlman et al. 2016, Brunsek et al. 2017). In addition, significant positive associations that studies do find tend to be modest—that is, effect sizes that roughly correspond to less than 0.10 standard deviations or 1.4 weeks of learning (Howes et al. 2008, Aikens et al. 2017).

The overall level of quality for higher-rated programs could not be described as high based on the program quality scores. For most states that used ERS as their measure of

program quality, the average scores for both higher- and lower-rated programs fell in the minimal range (scores of 3 or 4). This is considered better than inadequate (scores of 1 or 2), but is still worse than good (scores of 5 or 6) or excellent (a score of 7) (Harms et al. 2015). On the CLASS, in all but one of the states, both groups of programs had average scores for emotional support in the medium quality range (scores of 3 to 5). Both groups of programs had average scores in the low to medium (scores of 1-5) range on the instructional support scale (Pianta et al. 2008). These scores fall below the high quality range (scores of 6 or 7) on the CLASS.

Figure VI.1. Difference in program quality between higher- and lower-rated programs



Sources: State validation reports for California, Delaware, Massachusetts, Minnesota, Ohio, Oregon, Rhode Island, and Wisconsin.

Note: Washington is not shown in the figure because its validation report did not include analyses of program quality measures. Each black dot represents a state's difference in program quality (measured in standard deviations) for higher- versus lower-rated programs. Positive differences mean that higher-rated programs performed better than lower-rated programs. The blue dot represents the average difference across states. The lines on either side of the dots denote the 95 percent confidence interval. Two states calculated group means controlling for region (Wisconsin) and a variety of provider- and neighborhood-level characteristics, such as Title I status and the distribution of race in the provider's zip code (Delaware). The standardized mean differences using these regression-adjusted means are 0.64 standard deviations in Wisconsin (compared with 0.62 in the figure), and 0.26 standard deviations in Delaware (compared with 0.56 in the figure). The *p*-value for the Q test for heterogeneity is 0.01.

[a] The difference in program quality is calculated as the standardized mean difference, which is the difference in unadjusted means between higher- and lower-rated programs divided by the pooled standard deviation. The average is calculated using a statistical approach that weights each state's standardized mean difference by a measure of its confidence (a fixed-effect meta-analysis that weights states by their inverse variance).

VII. ASSOCIATIONS BETWEEN RATINGS AND CHILDREN'S DEVELOPMENT OUTCOMES

If programs with higher ratings are better at promoting children's development, children in higher-rated programs should have better outcomes than children in lower-rated programs. Eight states tested the validity of their TQRIS by examining the relationship between the ratings and measures of children's outcomes in domains that ranged from alphabetics to social-emotional development. Our analysis includes findings only when children had similar scores on child development measures at the beginning of the study, despite attending programs with different ratings; these findings are less likely to reflect differences between the types of children who attend different programs.[10]

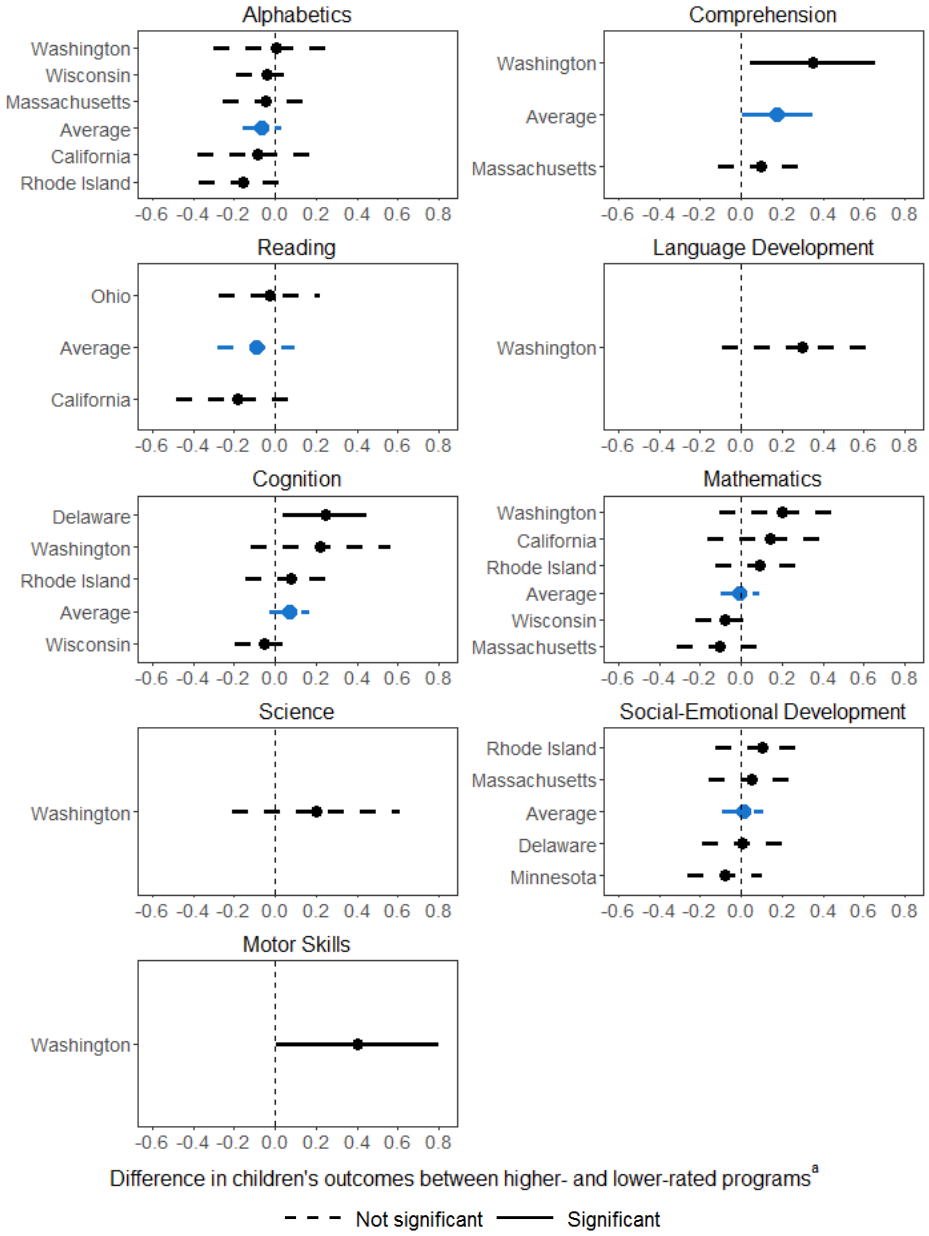
Outcomes for all children

In general, children attending higher-rated programs did not have better developmental outcomes than children attending lower-rated ones. Across the six child outcome domains that had findings for multiple states, children in higher-rated programs did not consistently score better, on average, than those in lower-rated programs across states (shown by the dashed 95 percent confidence intervals around the blue dots in Figure VII.1 that include zero). The average across states was significantly positive in only one domain: comprehension (shown by the solid 95 percent confidence interval around the blue dot in Figure VII.1 that does not include zero). In comprehension, children in higher-rated programs scored 0.18 standard deviations higher than children in lower-rated programs. Without regard to statistical significance, across domains, the average differences across states (shown by the blue dots in Figure VII.1) were both positive and negative.

Within each child outcome domain, few states saw significantly better scores for children in higher-rated programs, compared with those in lower-rated ones. For six child outcome domains, no states saw significantly better scores for children in higher-rated programs (as signified by the dashed 95 percent confidence intervals around the black dots in Figure VII.1 that include zero). For three domains—comprehension, cognition, and motor skills—only one state in each domain saw significantly higher scores for children in higher-rated programs (as signified by the solid 95 percent confidence intervals around the black dots in Figure VII.1 that do not include zero). The other states that analyzed comprehension and cognition did not find significant differences between children attending programs with different rating levels. No other state in the analysis assessed children's motor skills. Among domains that had findings for two or more states, findings did not vary significantly across states (based on a statistical test for heterogeneity called the "Q test").

[10] Sixty-nine percent (25 of 36) of the child outcome domains that states examined were based on children with similar scores at the beginning of the study and, thus, included in this analysis.

Figure VII.1. Differences in child development outcomes between children attending higher- and lower-rated programs, by domain



Sources: State validation reports for California, Delaware, Massachusetts, Minnesota, Ohio, Rhode Island, Washington, and Wisconsin.

Notes: Each black dot represents a state's effect size (measured in standard deviations) for higher- versus lower-rated programs. Positive effect sizes mean that children in higher-rated programs performed better than children in lower-rated programs. Each blue dot represents the average effect size across states. The lines on either side of the dots denote the 95 percent confidence interval. The *p*-value for the Q test for heterogeneity is 0.91 for alphabetics; 0.18 for comprehension; 0.43 for general reading achievement; 0.10 for cognition; 0.26 for mathematics; and 0.62 for social-emotional development.

[a] The difference in children's outcomes is an effect size (measured in standard deviations), which is calculated based on author-provided sample sizes, means, standard deviations, and other regression statistics. The average is calculated using a statistical approach that weights each state's effect size by a measure of its confidence (a fixed-effect meta-analysis that weights states by their inverse variance).

Most states did not consistently see higher scores for children in higher-rated programs, compared with those in lower-rated ones. Only Delaware and Washington saw significantly better scores for children attending higher-rated programs in at least one domain (Figure VII.2). Delaware found one statistically significant result in cognition but no significant result in social-emotional development. In Washington, only two outcomes across seven domains were statistically significant. No other states found any significant differences in outcomes. On the whole, there is no consistent evidence that children attending higher-rated programs scored higher on measures of child development, either across states or in any one state.

One hypothesis for why child outcomes did not differ for children attending higher- and lower-rated programs is because the two groups of programs might not have had sufficient differences in quality to lead to differences in children's learning. One might expect to see greater differences in children's outcomes if just the highest and lowest rating levels were compared.

To examine this hypothesis, we conducted a supplemental analysis that used findings from the highest and lowest individual rating levels (or groups of rating levels) that states reported. For example, if the state reported findings for five individual rating levels, the supplemental analysis compared level 1 (lowest) to level 5 (highest). To reduce burden on authors, we did not ask them to provide additional information for this analysis, so we used the statistics they reported to create the largest possible contrast in ratings (see Appendix B, Table B.1 for the comparisons used in each state). This analysis included the six states that reported information separately for each rating level, and excluded the two states that did not (because it was not possible to compare the highest and lowest rating levels in those states).

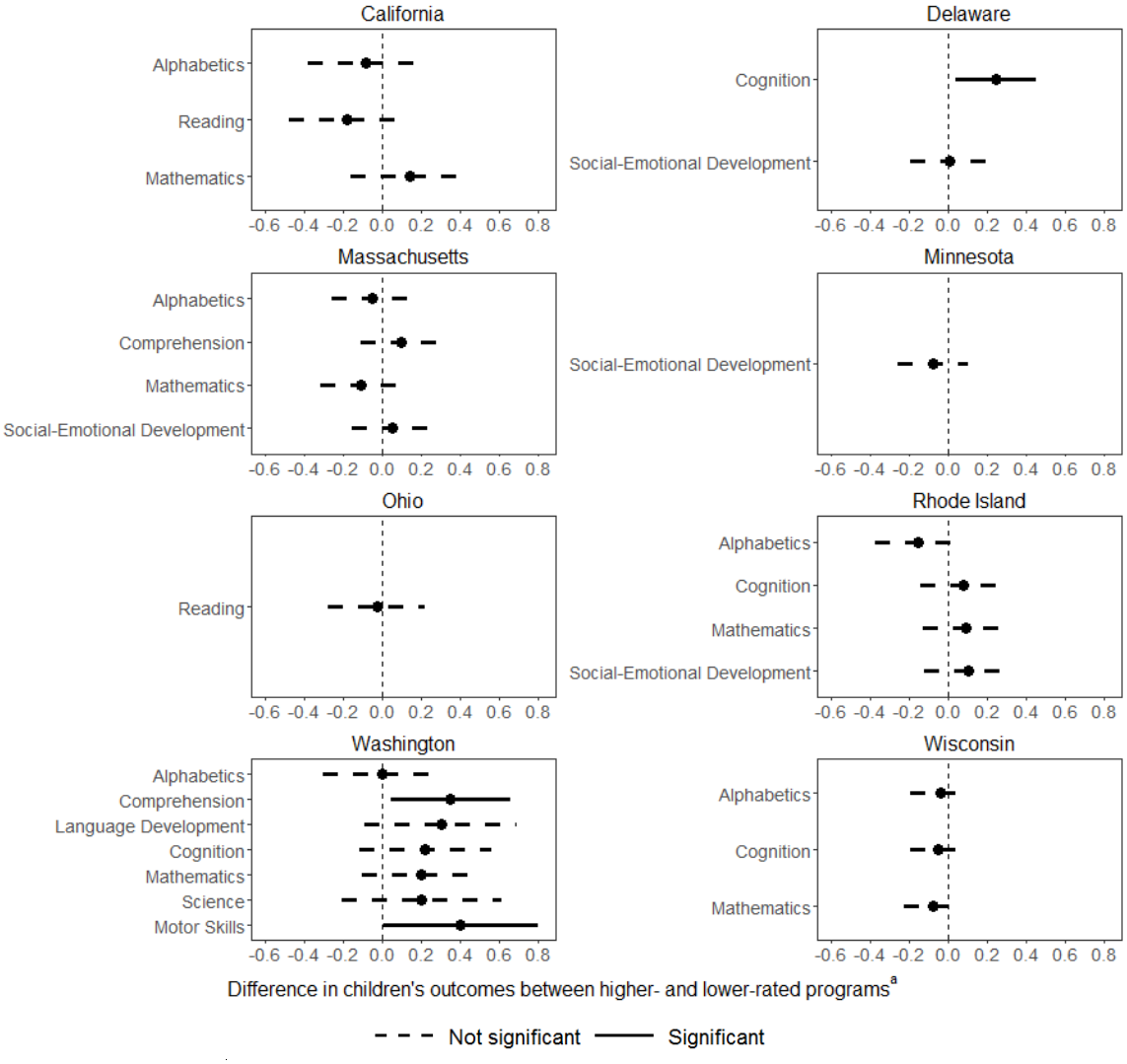
Even when comparing children from the highest- and lowest-rated programs, children attending higher-rated programs did not have better developmental outcomes than children attending lower-rated ones. For domains where there were enough states to perform the supplemental analysis, the estimated differences in outcomes between higher- and lower-rated programs increased slightly, but none were statistically significant (see Appendix B, Table B.2 for the results). Overall, the main analysis and this supplemental analysis provide no evidence that children attending higher-rated programs scored higher on measures of child development than children attending lower-rated programs.

Outcomes for low-income children

In addition to assessing the relationship between rating levels and child development outcomes overall, many states examined whether attending a higher-rated program was beneficial for children from low-income households.[11] High quality early childhood programs might be especially important for these children with fewer resources and opportunities to promote their development.

[11] The definitions of *low income* varied across the states. See Appendix C for state-specific definitions.

Figure VII.2. Difference in child development outcomes between children attending higher- and lower-rated programs, by state



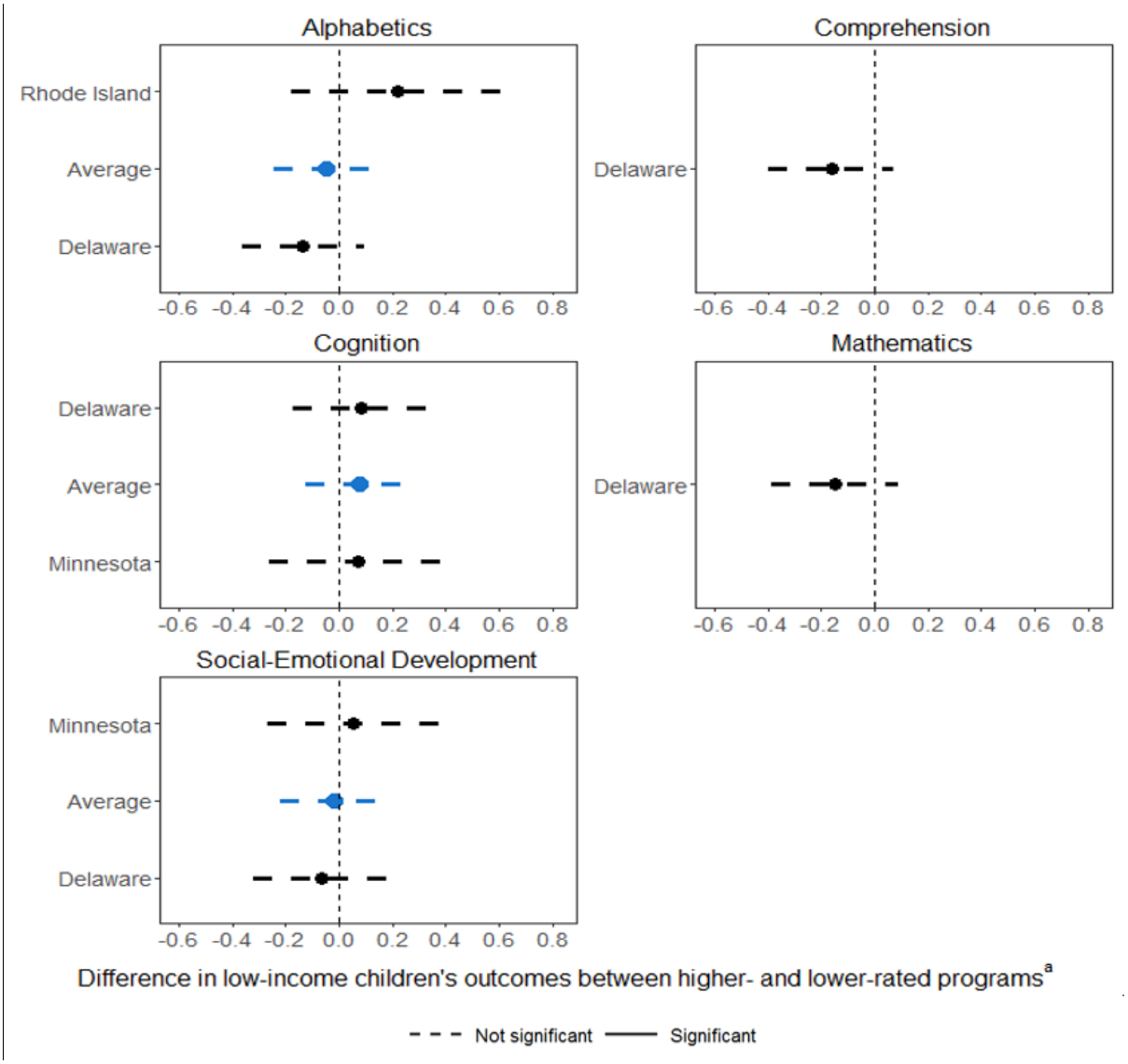
Sources: State validation reports for California, Delaware, Massachusetts, Minnesota, Ohio, Rhode Island, Washington, and Wisconsin.

Note: Each black dot represents a state's effect size (measured in standard deviations) for higher- versus lower-rated programs. Positive effect sizes mean that children in higher-rated programs performed better than children in lower-rated programs. Each blue dot represents the average effect size across states. The lines on either side of the dots denote the 95 percent confidence interval.

[a] The difference in children's outcomes is an effect size (measured in standard deviations), which is calculated based on author-provided sample sizes, means, standard deviations, and other regression statistics.

Ratings of TQRIS also were not associated with child development outcomes for low-income children. In the five domains in which states analyzed outcomes separately for low-income children—alphabetics, comprehension, cognition, mathematics, and social-emotional development—none of the states found significant differences between the scores of low-income children in higher- and lower-rated programs on any outcome (Figure VII.3).

Figure VII.3. Differences in child development outcomes between low-income children attending higher- and lower-rated programs, by domain



Sources: State validation reports for Delaware, Minnesota, and Rhode Island.

Note: Each black dot represents a state's effect size (measured in standard deviations) for higher- versus lower-rated programs. Positive effect sizes mean that children in higher-rated programs performed better than children in lower-rated programs. Each blue dot represents the average effect size across states. The lines on either side of the dots denote the 95 percent confidence interval.

[a] The difference in children's outcomes is an effect size (measured in standard deviations), which is calculated based on author-provided sample sizes, means, standard deviations, and other regression statistics. The average is calculated using a statistical approach that weights each state's effect size by a measure of its confidence (a fixed-effect meta-analysis that weights states by their inverse variance).

Potential explanations for findings for all children and low-income children

Overall, there is no consistent evidence that children who participated in programs with higher rating levels had better outcomes, overall or specifically among low-income children. There are several potential explanations for this finding.

It is unlikely that there would have been larger differences in outcomes if all of the validation studies had compared only children in programs with the lowest rating level to children in programs with the highest rating level. Such an analysis would provide the largest differences in program quality with which to best assess relationships with developmental outcomes. We conducted this analysis to the extent possible with the subset of states that compared rating levels individually, and we found no statistically significant differences across states in any domain. In the validation studies for the seven states that did compare rating levels individually (California, Delaware, Massachusetts, Ohio, Oregon, Rhode Island, and Washington), there were either no statistically significant differences or sporadic differences (most of which were in the hypothesized direction but some of which were not).

It is unlikely that the amount of time between baseline and follow-up assessments was insufficient for detecting changes in children's skills. It is possible that a longer follow-up period could have resulted in larger differences in children's outcomes. Most states conducted baseline assessments in the fall and follow-up assessments in the spring, so there was a range of about one to nine months between assessments. Nearly all states showed that children made age-appropriate gains between baseline and follow-up on the child development measures they collected and analyzed for their studies. Therefore, the time between assessments was sufficient to detect gains, but those gains were not greater for children in higher-rated programs.

The differences in program quality between higher- and lower-rated programs might not be large enough to generate differences in child development outcomes. Across states, the average difference in program quality corresponded to roughly half a point on the ERS or CLASS, which might not be large enough to cause differences in children's development. For example, one study found that a point difference on the ERS or CLASS emotional support score was not associated with differences in children's vocabulary, alphabetics, or math skills, but a point difference on CLASS instructional support was (Mashburn et al. 2008).

The states that saw the largest differences in program quality (California and Rhode Island) did not see significant differences in any of the child outcome domains. Washington found the greatest number of significant differences in children's outcomes, but did not analyze an external measure of program quality.

The states' validation studies also did not find strong evidence of an association between the program quality measures and children's outcomes. All five states that reported on these associations found few or no positive associations. The findings are consistent with previous literature that finds, at best, modest positive associations (that is, effect sizes of less than 0.10 standard deviations) between program quality measures and children's outcomes (Howes et al. 2008; Perlman et al. 2016; Brunsek et al. 2017).

The average levels of quality among programs in TQRIS were not high enough to affect children's outcomes. As previously mentioned, the ERS and CLASS scores of high- and

low-rated programs were not in the range described as high quality by the publishers of these measures. Recent studies suggest that there might be a particular threshold at which differences in quality result in differences in children's outcomes (Burchinal et al. 2016; Weiland et al. 2013).

The ways that states calculate and award ratings could weaken the relationship between these ratings and children's development. Using a large number of indicators to determine the ratings (including some that are not closely related to children's classroom experiences) might contribute to a lack of correspondence between these ratings and children's development. Previous studies have found stronger relationships between more specific measures of quality compared with broader measures. For example, one study found that measures of the quality of instruction in literacy or math or of teacher–child interactions were stronger predictors of children's outcomes than broader measures of the quality of the classroom environment (Burchinal et al. 2016). Another study found that measures of the quality of teacher–child interactions had stronger relationships to children's outcomes than simulated ratings (Sabol et al. 2013). In fact, only a few of the components of the ratings (such as child:staff ratios and group size, curriculum, staff qualifications, and quality of the environment) have been associated with children's outcomes (Kirby et al. 2017).

The rating structures and state policies, such as automatic ratings, might also weaken the relationship between ratings and children's development. In block rating structures, programs that just miss qualifying for the next rating level might not have sufficiently different child outcomes from programs that barely achieved the next rating level. In hybrid and points rating structures, programs could reach rating levels in different ways, and not all of these might be strongly associated with children's development. Automatic ratings could also weaken the relationship between children's outcomes if programs that automatically received top ratings would not have received those ratings through the full data collection and verification process.

The analyses might have lacked sufficient statistical power to detect differences in outcomes between higher- and lower-rated programs. Having a larger number of programs could improve the analyses' ability to detect statistically significant differences between programs. However, this does not explain the patterns of small differences in some domains and negative differences in others. Some states also found a few statistically significant differences with the number of programs in their analyses, but other states' analyses could have lacked sufficient statistical power.

VIII. CHALLENGES IN CONDUCTING VALIDATION STUDIES

The researchers who conducted the validation studies reported challenges in four broad areas: study design, sample size, measure selection and data collection, and the schedule and timing of the study. We discuss common issues that at least five researchers perceived as either a major or minor challenge. See Appendix D for complete findings on all of the challenges that we discussed with researchers in the interviews.

All nine researchers noted at least one limitation in their validation study design or analysis approach that could affect the interpretation of findings. Researchers often cited having a non-experimental design (in which families choose child care programs rather than being randomly assigned to them) as a limitation (Table VIII.1). In non-experimental designs, differences between the outcomes of children in higher- versus lower-rated programs might reflect differences between the characteristics of families that choose these programs, rather than true differences due to programs' quality. Ultimately, few researchers (three of nine) perceived the design limitations as major; most described them as only minor challenges requiring a careful approach to developing research questions and interpreting findings.

Table VIII.1. Common challenges reported by researchers who conducted validation studies

Challenge	Number of states categorizing as a challenge (major or minor)	Number of states categorizing as a major challenge
Study design and analysis approach		
Design limited the interpretation of findings	9	3
Deciding which rating levels to validate	5	1
Sample size and representativeness of sample		
Recruiting programs	8	4
Attaining sufficient representation across program types and rating levels	8	5
Recruiting children	5	1
Low response rates from programs	5	0
Measures and collecting necessary data		
Selecting measures of child development	7	2
Missing data for programs or children	7	1
Analyzing administrative data[a]	6	1
Analyzing measures of child development	5	1
Obtaining administrative data[a]	5	2
Selecting measures of program quality	5	0
Limited data on family/child characteristics	5	2
Schedule and timing of study		
Conducting study before TQRIS were fully implemented[b]	6	4
Collecting data in the allotted study time frame	6	1

Sources: Interviews with principal investigators who conducted the state validation reports in California, Delaware, Massachusetts, Minnesota, Ohio, Oregon, Rhode Island, Washington, and Wisconsin.

Note: Table includes only challenges that at least five states reported.

[a] Applies to only eight states.

[b] This issue applied to only a subset of states. TQRIS were not fully implemented or were in a major transition period at the start of the validation study in six of the nine states.

The most common challenges related to sample size and representativeness were recruiting programs for the study and attaining sufficient representation across rating levels or program types. Eight of the nine researchers cited each of these as challenges. Recruiting child care providers for the study (categorized as major by four of nine) and attaining sufficient representation across rating levels or provider types (categorized as major by five of nine) were perceived as particularly challenging. Those two challenges were categorized as major more frequently than any other (Table VIII.1). For example, researchers in four states noted that recruiting family child care providers was particularly difficult. As one researcher pointed out, family child care providers are typically smaller operations with fewer resources to offset the burden of participating in a study. Family child care providers and the families they serve also tend to be less stable and harder to reach than center-based programs. Researchers most often struggled to recruit a sufficient number of child care providers from each rating level (five of eight reported recruiting sample challenges). This challenge typically resulted from having very few child care providers at certain rating levels, most often the highest levels.

The most common challenges related to measuring child development outcomes and collecting data were selecting measures and having missing data for programs or children. Seven of the nine researchers cited each of these as challenges. Researchers for six validation studies reported challenges selecting child development measures, noting the difficulty in striking a balance between avoiding an excessively lengthy battery of tests and measuring a broad mix of outcomes. On average across the eight states that collected child development data, a state administered six distinct child assessments (each of which might have included multiple subscales). The number of child assessments administered by each state ranged from 2 to 12. Two of the six researchers also encountered challenges sourcing a test battery available in multiple languages. Researchers also struggled with missing data; two researchers specifically cited missing values or variables in administrative data as problematic, whereas four researchers faced challenges with missing data for families and children, due to both attrition and low response rates on parent surveys.

In six states where the validation study began before or concurrent with full implementation of TQRIS, researchers viewed this timing for conducting the validation study as problematic. When TQRIS were not fully implemented at the time of the study, researchers experienced challenges identifying eligible centers and their rating categories, given that ratings could be outdated. Further, providers that were confused or overburdened by system changes were frequently reluctant to participate and difficult to recruit. Findings based on systems in flux might also be difficult to interpret.

Researchers also reported additional challenges, not specifically queried in the interviews, about relating the ratings to children's outcomes. First, researchers for two studies noted that research questions related to children's outcomes, which are often distal from child care program quality, were unrealistic given the timeline for the validation studies. They emphasized the need for a more longitudinal perspective when examining the relationship between the rating levels and kindergarten readiness or other outcomes. Two other researchers pointed to characteristics of the system that might result in ratings not accurately reflecting program quality. One noted that programs can choose to stay at the same rating level if there is not sufficient incentive to achieve a higher rating. Another researcher pointed out that the ratings can mask variation among programs: at a particular level, programs can exceed at least some of the standards for their rating.

IX. DISCUSSION

This synthesis of validation studies conducted by nine RTT-ELC states provides some evidence that ratings of TQRIS capture differences in program quality. Compared with lower-rated programs, programs with higher rating levels had significantly higher scores on independently collected measures of program quality in nearly all states. However, differences in program quality reflected in the ratings did not translate into differences in children's outcomes. This finding is largely consistent with the findings of previous validation studies, the findings of the simulation conducted by Sabol and colleagues (2013), and a recent synthesis of state studies conducted by Tout et al. (2017).

There are several likely reasons children in higher-rated programs did not consistently perform better than children in lower-rated programs. First, the overall levels of quality among programs, especially among higher-rated programs in TQRIS, were not high by publishers' standards. Second, the differences in program quality might not have been large enough to produce significant differences in children's outcomes. Third, rating levels might not align specifically enough to practices that would influence children's development. The quality of classroom interactions and instructional practices is typically only one of numerous individual components upon which the ratings are based. Furthermore, TQRIS are structured to promote participation and quality improvement by giving programs flexibility in how they obtain ratings. Some of this flexibility comes in the form of points or hybrid systems or in the availability of automatic rating options. These characteristics might contribute to the lack of a relationship because the rating levels capture quality in a broad and inconsistent way.

The program quality and child development findings underscore a challenge in implementing TQRIS to meet multiple goals. States design their systems to draw attention to multiple dimensions of quality that are important (Kirby et al. 2015; Zellman and Perlman 2008). However, some dimensions, such as the quality of program administration, are not as closely related to children's experiences as what happens in the classroom. How much this matters for the ultimate goals of TQRIS may depend on how the ratings are used. Ratings that are not associated with children's outcomes could still help states identify and support programs that are low-performing on other important dimensions. Yet, if the key objective of TQRIS is to help states and parents identify programs that improve children's developmental outcomes, states may want to consider this objective when making revisions to their TQRIS design.

As TQRIS become more fully implemented, it will be necessary to conduct additional studies that examine the validity of ratings. This synthesis documented that researchers faced challenges in designing and executing studies to answer the questions of interest related to the relationships between the ratings and program quality and children's outcomes given the implementation status of TQRIS and the time frames in which they conducted these studies. When TQRIS are first rolled out, it can be important for validation studies to assess whether the measures and ratings are implemented as planned, in addition to examining the relationships between the ratings and external quality measures or assessments of children's developmental outcomes (Tout and Starr 2013).

As TQRIS continue to mature and include more programs, the goals of future validation studies could change. As Zellman and Fiene (2012) have pointed out, validation should be an

ongoing effort that not only sheds light on whether ratings are useful, but also helps inform refinements to improve the system.

Although our findings were relatively consistent across states, several limitations hamper our ability to draw conclusions from the synthesis of state validation studies. First, our synthesis is limited to only nine states with wide variability in TQRIS. Second, this synthesis is based on non-experimental designs that might not completely account for existing differences between children attending higher- and lower-rated programs. We also relied on the analyses presented by the researchers in each state. This meant that we could not compare individual rating levels in all states because not all states did so. We also did not have data for individual programs or on each component, which could be used to better understand why programs with different rating levels did not have larger differences in children's outcomes. Along similar lines, the analyses that researchers presented focused primarily on preschoolers in center-based programs. Some researchers combined data on preschoolers with data on infants and toddlers, or combined data for centers with data for family child care programs, so we did not have sufficient data to look at whether ratings were more or less predictive based on children's age or based on the type of program. Finally, the responses researchers gave in the interviews were relative to each researcher's previous experience. Thus, the extent to which they viewed particular elements as challenging likely varied for reasons outside of the validation study itself. Additional states will release validation reports that will contribute to our knowledge about whether TQRIS differentiate between programs of differing quality and whether those differences are related to children's development.

REFERENCES

Aikens, N., A. Kopack Klein, E. Knas, J. Hartog, M. Manley, L. Malone, L. Tarullo, and S. Lukashanets. "Child and Family Outcomes During the Head Start Year: FACES 2014–2015 Data Tables and Study Design." OPRE Report #2017-100. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2017.

Arnett, J. "Caregivers in Day Care Centers: Does Training Matter?" *Journal of Applied Developmental Psychology,* vol. 10, no. 4, 1989, pp. 541–552.

Brunsek, A., M. Perlman, O. Falenchuk, E. McMullen, B. Fletcher, and P.S. Shah. "The Relationship Between the Early Childhood Environment Rating Scale and Its Revised Form and Child Outcomes: A Systematic Review and Meta-Analysis." *PLoS ONE,* vol. 12, no. 6, 2017.

Bryant, D.M., K. Bernier, K. Maxwell, and E.S. Peisner-Feinberg. "Validating North Carolina's 5-Star Child Care Licensing System." Chapel Hill, NC: University of North Carolina, Frank Porter Graham Child Development Center, 2001.

Build Initiative and Child Trends. "A Catalog and Comparison of Quality Rating and Improvement Systems (QRIS)." Data system. 2016. Available at http://qriscompendium.org/. Accessed November 12, 2017.

Burchinal, M.R., Roberts, J.E., Nabors, L.A., and Bryant, D.M. "Quality of Center Child Care and Infant Cognitive and Language Development." *Child Development,* vol. 67, no. 2, 1996, pp. 606–620.

Burchinal, M., L. Tarullo, and M. Zaslow. "Best Practices in Creating and Adapting Quality Rating and Improvement System (QRIS) Rating Scales." OPRE Research Brief #2016-25. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2016.

Burchinal, M., Y. Xue, A. Auger, H. Tien, A. Mashburn, E. Peisner-Feinberg, E. Cavadel, M. Zaslow, and L. Tarullo. "III. Testing for Quality Thresholds and Features in Early Care and Education." *Monographs of the Society for Research in Child Development,* vol. 81, no. 2, 2016, pp. 46–63.

Clifford, R.M., S.S. Reszka, and H-G. Rossbach. "Reliability and Validity of the Early Childhood Environment Rating Scale." Working paper. Chapel Hill, NC: University of North Carolina at Chapel Hill, January 2010.

Dearing, E., K. McCartney, and B.A. Taylor. "Does Higher Quality Early Child Care Promote Low-Income Children's Math and Reading Achievement in Middle Childhood?" *Child Development,* vol. 80, 2009, pp. 1329–1349.

Elicker, J., C.C. Langill, K.M. Ruprecht, J. Lewsader, and T. Anderson. "Evaluation of Paths to QUALITY, Indiana's Child Care Quality Rating and Improvement System: Final Report." West Lafayette, IN: Purdue University, 2011.

Gordon, R.A., K. Fujimoto, R. Kaestner, S. Korenman, and K. Abner. "An Assessment of the Validity of the ECERS-R with Implications for Assessments of Child Care Quality and Its Relation to Child Development." *Developmental Psychology,* vol. 49, issue 1, 2013, pp. 146–160. doi:10.1037/a0027899.

Gordon, R.A., K.G. Hofer, K. A. Fujimoto, N. Risk, R. Kaestner, and S. Korenman. "Identifying High-Quality Preschool Programs: New Evidence on the Validity of the Early Childhood Environment Rating Scale–Revised (ECERS-R) in Relation to School Readiness Goals." *Early Education and Development,* vol. 26, issue 8, 2015, pp. 1086–1110. doi: 10.1080/10409289.2015.1036348.

Harms, T., D. Cryer, and R. Clifford. *Infant/Toddler Environment Rating Scale - Revised Edition.* New York: Teachers College Press, 2003.

Harms, T., R. Clifford, and D. Cryer. *Early Childhood Environment Rating Scale-Revised Edition—Updated*. New York: Teachers College Press, 2005.

Harms, T., D. Cryer, and R. Clifford. *Family Child Care Environment Rating Scale-Revised Edition*. New York: Teachers College Press, 2007.

Heinemeier, S., A. D'Agostino, J. Hamilton, K. Kim, and M. Winglee. "Ohio's SUTQ Validation Study Results." Durham, NC: Compass Evaluation and Research, Inc., 2017.

Howes, C., M. Burchinal, R. Pianta, D. Bryant, D. Early, R. Clifford, and O. Barbarin. "Ready to Learn? Children's Pre-Academic Achievement in Pre-Kindergarten Programs." *Early Childhood Research Quarterly,* vol. 23, issue 1, 2008, pp. 27–30.

Karoly, L.A., H.L. Schwartz, C. Messan Setodji, and A.C. Haas. "Evaluation of Delaware Stars for Early Success: Final Report." Santa Monica, CA: RAND Corporation, 2016.

Kirby, G., P. Caronongan, L.M. Malone, and K. Boller. "What Do Quality Rating Levels Mean? Examining the Implementation of QRIS Ratings to Inform Validation." *Early Childhood Research Quarterly,* vol. 30, part B, 2015, pp. 91–305.

Kirby, G., P. Caronongan, A. Mraz Esposito, L. Murphy, M. Shoji, P. Del Grosso, W. Kiambuthi, M. Clark, and L. Dragoset. "Progress and Challenges in Developing Tiered Quality Rating and Improvement Systems (TQRIS) in the Round 1 Race to the Top-Early Learning Challenge (RTT-ELC) States." NCEE 2018-4003. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2017.

Lahti, M., C. Cobo-Lewis, A. Dean, S. Rawlings, E. Sawyer, and B. Zollitsch. "Maine's Quality for Me: Child Care Quality Rating and Improvement System (QRIS): Final Evaluation Report." Augusta, ME: Department of Health and Human Services, 2011.

Lahti, M., T. Sabol, R. Starr, C. Langill, and K. Tout. "Validation of Quality Rating and Improvement Systems: Examples from Four States." Research-to-Policy, Research-to-Practice Brief OPRE 2013-036. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2013.

Lipscomb, S.T., R.B. Weber, B.L. Green, and L.B. Patterson. "Oregon's Quality Rating Improvement System (QRIS) Validation Study One: Associations with Observed Program Quality." San Mateo, CA: American Institutes for Research, 2016.

Loeb, S., B. Fuller, S.L. Kagan, and B. Carrol. "Child Care in Poor Communities: Early Learning Effects of Type, Quality, and Stability." *Child Development,* vol. 75, 2004, pp. 47-65.

Magnuson, K., and Y.C. Lin. "Validation of QRIS YoungStar Rating Scale. Report 1." Madison, WI: School of Social Work and Institute for Research on Poverty, 2015.

Magnuson, K., and Y.C. Lin. "Validation of QRIS YoungStar Rating Scale. Report 2." Madison, WI: School of Social Work and Institute of Research on Poverty, 2016.

Malone, L., G. Kirby, P. Caronongan, K. Boller, and K. Tout. "Measuring Quality Across Three Child Care Quality and Improvement Systems: Findings from Secondary Analyses." OPRE Report #2011-30. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2011.

Mashburn, A.J., R.C. Pianta, B.K. Hamre, J.T. Downer, O.A. Barbarin, D. Bryant, M. Burchinal, D.M. Early, and C. Howes. "Measures of Classroom Quality in Prekindergarten and Children's Development on Academic, Language, and Social Skills." *Child Development,* vol. 79, no. 3, 2008, pp. 732–749.

Maxwell, K.L., A. Blasberg, D.M. Early, and N. Orfali. "Evaluation of Rhode Island's BrightStars Child Care Center and Preschool Quality Framework." Chapel Hill, NC: Child Trends, 2016.

Norris, D.J., and L. Dunn. "Reaching for the Stars: Family Child Care Home Validation Study Final Report." Stillwater and Norman, OK: Early Childhood Collaborative of Oklahoma, 2004.

Perlman M., O. Falenchuk, B. Fletcher, E. McMullen, J. Beyene, and P.S. Shah. "A Systematic Review and Meta-Analysis of a Measure of Staff/Child Interaction Quality (the Classroom Assessment Scoring System) in Early Childhood Education and Care Settings and Child Outcomes." *PLOS One,* 2016. Available at https://doi.org/10.1371/journal.pone.0167660.

Pianta, R., K. LaParo, and B. Hamre. "The Classroom Assessment Scoring System Pre-K Manual." Charlottesville, VA: University of Virginia, 2008.

Quick, H.E., L.E. Hawkinson, A. Holod, J. Anthony, S. Muenchow, D. Parrish, A. Martin, E. Weinberg, D.H. Lee, J.S. Cannon, L.A. Karoly, G.L. Zellman, S. Faxon-Mills, A. Muchow, and T. Tsai. "Independent Evaluation of California's Race to the Top-Early Learning Challenge: Cumulative Technical Report." San Mateo, CA: American Institutes for Research, 2016a.

Quick, H.E., A. Holod, D.H. Lee, E. Weinberg, S. Muenchow, and L.E. Hawkinson. "Independent Evaluation of California's Race to the Top-Early Learning Challenge: Supplemental Validation Study Report." San Mateo, CA: American Institutes for Research, 2016b.

Roberts, J., N.L. Marshall, A. Tracy, S. Santaniello, M.G. Melia, H. Moore, S. Ellis, L. Kaufman, J. Tapper, S. Leibowitz, B. Comer, K. Gibbons, A. Glasgow, and K. Khlifi. "Massachusetts Quality Rating and Improvement System (QRIS) Validation Study Final Report." Wellesley, MA: Wellesley Centers for Women, 2016.

Sabol, T.J., and R.C. Pianta. "Improving Child Care Quality: A Validation Study of the Virginia Star Quality Initiative." Charlottesville, VA: Center for the Advanced Study of Teaching and Learning, 2012.

Sabol, T.J., S.L. Soliday Hong, R.C. Pianta, and M.R. Burchinal. "Can Rating Pre-K Programs Predict Children's Learning?" *Science,* vol. 341, 2013, pp. 845–846.

Soderberg, J., G.E. Joseph, S. Stull, and N. Hassairi. "Early Achievers Standards Validation Study." Seattle, WA: Childcare Quality and Early Learning Center for Research & Professional Development, 2016.

Teachstone. "Effective Teacher-Child Interactions and Child Outcomes: A Summary of Research on the Classroom Assessment Scoring System (CLASS) Pre-K–3rd Grade." Charlottesville, VA: Teachstone, 2017.

Thornburg, K., W. Mayfield, J. Hawks, and K. Fuger. "The Missouri Quality Rating System School Readiness Study." Columbia, MO: University of Missouri Center for Family Policy and Research, 2009.

Tout, K., R. Starr, T. Isner, J. Cleveland, L. Albertson-Junkans, M. Soli, and K. Quinn. "Evaluation of Parent Aware: Minnesota's Quality Rating and Improvement System Pilot: Final Evaluation Report." St. Paul, MN: Minnesota Early Learning Foundation, 2011.

Tout, K., J. Cleveland, W. Li, R. Starr, M. Soli, and E. Bultinck. "The Parent Aware Evaluation: Initial Validation Report." Minneapolis, MN: Child Trends, 2016.

Tout, K., and R. Starr. "Key Elements of a QRIS Validation Plan: Guidance and Planning Template." OPRE Report #2013-11. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2017.

Tout, K., K. Magnuson, S. Lipscomb, L. Karoly, R. Starr, H. Quick, D. Early, D. Epstein, G. Joseph, K. Maxwell, J. Roberts, C. Swanson, and J. Wenner. "Validation of the Quality Ratings Used in Quality Rating and Improvement Systems (QRIS): A Synthesis of State Studies." OPRE Report #2017-92. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2017.

U.S. Department of Education. "Race to the Top-Early Learning Challenge FY 2013 Competition: Guidance and Frequently Asked Questions for Applicants." Washington, DC: U.S. Department of Education, 2013. Available at https://www2.ed.gov/programs/racetothetop-earlylearningchallenge/faqs-9-13-2013.pdf. Accessed November 12, 2017.

Weiland, C., K. Ulvestad, J. Sachs, and H. Yoshikawa. "Associations Between Classroom Quality and Children's Vocabulary and Executive Function Skills in an Urban Public Prekindergarten Program." *Early Childhood Research Quarterly,* vol. 28, no. 2, 2013, pp. 199–209.

U.S. Department of Education. *What Works Clearinghouse Standards Handbook Version 4.0.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2017.

U.S. Department of Education. *What Works Clearinghouse Procedures and Standards Handbook Version 3.0.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, 2014.

Yoshikawa, H., C. Weiland, J. Brooks-Gunn, M.R. Burchinal, L.M. Espinosa, W.T. Gormley, J. Ludwig, K.A. Magnuson, D. Phillips, and M.J. Zaslow. "Investing in Our Future: The Evidence Base on Preschool Education." Washington, DC: Society for Research in Child Development and Foundation for Child Development, 2013.

Zellman, G.L., M. Perlman, V. Le, and C.M. Setodji. "Assessing the Validity of the Qualistar Early Learning Quality Rating and Improvement System as a Tool for Improving Child Care Quality." Santa Monica, CA: RAND Corporation, 2008.

Zellman, G.L., and M. Perlman. "Child-Care Quality Rating and Improvement Systems in Five Pioneer States: Implementation Issues and Lessons Learned." Santa Monica, CA: RAND Corporation, 2008.

Zellman, G.L., and R. Fiene. "Validation of Quality Rating and Improvement Systems for Early Care and Education and School-Age Care." Research-to-Policy, Research-to-Practice Brief OPRE 2012-29. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2012.

APPENDIX A

OUTCOME DOMAINS AND MEASURES

This appendix lists all of the outcome domains and measures that states used in their validation studies and provides more information about measures of program quality. Table A.1 shows the measures used in each domain. Table A.2 shows the measures used by each state. Table A.3 describes two common measures of program quality: (1) the Early Childhood Environment Rating Scale (ECERS), and (2) the Classroom Assessment Scoring System (CLASS).

Table A.1. Outcome domains used for establishing validity

Outcome domain	Description	Measures used in state validation reports
Cognition	Includes outcomes in the following areas: memory, problem-solving, cognitive processing and flexibility, and general knowledge (including school readiness and intelligence quotient [IQ])	• Bracken School Readiness Assessment (BSRA) • Peg Tapping • Head-Toes-Knees-Shoulders (HTKS) • Pencil Tap Test • Mullen Scales of Early Learning (MSEL), Visual Reception
Mathematics	Includes outcomes in the following areas: basic number concepts, number operations, patterns and classification, measurement, geometry, and general numeracy	• Woodcock Johnson Tests of Achievement (WJ III), Applied Problems • Tools for Early Assessment in Math (TEAM)
Language development	Includes outcomes that assess the ability to understand spoken language, communicate and understand thoughts or ideas through speech, use developmentally appropriate discourse skills, and display grammatical knowledge or skill	• Mullen Scales of Early Learning (MSEL), Expressive Language and Receptive Language • Brigance Inventory of Early Development (IED), Language development
Alphabetics	Includes outcomes in the following areas: phonemic and phonological awareness, letter identification, print awareness, and phonics	• Woodcock Johnson Tests of Achievement (WJ III), Letter-Word Identification • Test of Preschool Early Literacy (TOPEL) • Early Writing Assessment (EWA)
Comprehension	Includes outcomes in the areas of vocabulary and comprehension development	• Woodcock Johnson Tests of Achievement (WJ III), Picture Vocabulary • Peabody Picture Vocabulary Test (PPVT) • Individual Growth and Developing Indicators (IGDI)
General reading achievement	Includes outcomes that combine measures in two or more of the previous domains (for example, alphabetics and comprehension) or provide some other type of summary score across domains, such as a total reading score on a standardized reading test	• Story and Print Concepts • Brigance Inventory of Early Development (IED), Literacy
Science	Includes outcomes related to children's content and processing skill knowledge in science	• Lens on Science (LENS)
Social-emotional development	Includes outcomes in the following areas: behavioral, social, and emotional competencies underlying school readiness, such as pro-social (or problem) behaviors, social interactions, cooperation, self-concept, engagement, attention, persistence, impulsivity, self-control, and initiative	• Devereaux Early Childhood Assessment (DECA) • Preschool Learning Behaviors Scale (PLBS) • Social Competence and Behavior Evaluation (SCBE) • Child Behavior Checklist (CBCL)
Motor skills	Includes outcomes measuring either fine and/or gross motor skills	• Mullen Scales of Early Learning (MSEL), Fine Motor and Gross Motor
Program quality	Includes program-level measures of observed quality and curriculum practices	• Early Childhood Environment Rating Scale-Revised (ECERS-R) • Early Childhood Environment Rating Scale-Revised (ECERS-3) • Family Child Care Environment Rating Scale-Revised (FCCERS-R) • Infant/Toddler Environment Rating Scale-Revised (ITERS-R) • Classroom Assessment Scoring System (CLASS) • Early Language and Literacy Classroom Observation (ELLCO) • Child Home Early Language and Literacy Observation (CHELLO) • Arnett Caregiver Interaction Scale (CIS) • Preschool Program Quality Assessment (PQA)

Sources: State validation reports for California, Delaware, Massachusetts, Minnesota, Ohio, Oregon, Rhode Island, Washington, and Wisconsin.

Table A.2. Outcome measures used for establishing validity

Outcome measure	CA	DE	MA	MN	OH	OR	RI	WA	WI
Alphabetics									
WJ III Letter-Word Identification	X	X	X				X	X	X
TOPEL				X					X
EWA								X	
Comprehension									
WJ III Picture Vocabulary							X		
PPVT		X	X					X	
IGDI				X					
General reading achievement									
Story and Print Concepts	X								
Brigance (IED) Literacy					X				
Language development									
MSEL Expressive Language								X	
MSEL Receptive Language								X	
Brigance (IED) Language development					X				
Cognition									
BSRA									X
Peg Tapping	X			X					
HTKS		X						X	X
Pencil Tap Test							X		
MSEL Visual Reception								X	
Mathematics									
WJ III Applied Problems	X	X	X	X			X		X
TEAM								X	
Science									
LENS								X	
Social-emotional development									
DECA		X	X						
PLBS			X	X					X
SCBE				X			X		X
CBCL								X	
Motor skills									
MSEL Fine Motor								X	
MSEL Gross Motor								X	
Program quality									
ECERS-R			X	X					X
ECERS-3					X				
FCCERS-R				X	X				X
ITERS-R			X		X				
CLASS	X	X		X	X	X	X		
ELLCO					X				
CHELLO					X				
CIS		X	X		X				
PQA		X							

Sources: State validation reports for California, Delaware, Massachusetts, Minnesota, Ohio, Oregon, Rhode Island, Washington, and Wisconsin.

Table A.3. Common program quality outcome measures

Program quality measure	Description	Domains	Scoring	Evidence of predictive validity
Early Childhood Environment Rating Scale-Revised (ECERS-R)	Instrument that measures quality of classroom environment	Space and Furnishings, Personal Care Routines, Language Reasoning, Activities, Interaction, Program Structure, and Parents and Staff	1-2 Inadequate 3-4 Minimal 5-6 Good 7 Excellent	Studies find significant positive associations between ECERS-R scores or subscores and children's cognition or social outcomes (Clifford et al. 2010).
Family Child Care Environment Rating Scale-Revised (FCCERS-R)	Instrument that measures quality of classroom environment	Space and Furnishings, Personal Care Routines, Listening and Talking, Activities, Interaction, Program Structure, Parents and Provider	1-2 Inadequate 3-4 Minimal 5-6 Good 7 Excellent	The scale is part of environmental rating scales that are predictive of child outcomes (Harms et al. 2007).
Infant/Toddler Environment Rating Scale Revised (ITERS-R)	Instrument that measures quality of classroom environment	Space and Furnishings, Person Care Routines, Listening and Talking, Activities, Interaction, Program Structure, Parents and Staff	1-2 Inadequate 3-4 Minimal 5-6 Good 7 Excellent	Studies find significant positive associations between ITERS and cognitive development, language development, and communication skills (Burchinal et al. 1996).
Classroom Assessment Scoring System (CLASS)	Instrument that measures teacher-child interactions in classroom	Emotional Support, Classroom Organization, and Instructional Support	1-2 Low 3-5 Medium 6-7 High	Studies find significant positive associations between CLASS scores or domain scores and children's academic or social-emotional outcomes (Teachstone 2017).
Caregiver Interaction Scale (CIS)	Instrument that measures teacher-child interactions in classroom	Sensitivity, Harshness, Detachment, Permissiveness	1 Not true at all 2 Somewhat true 3 Quite a bit true 4 Very much true	Studies find significant positive associations between CIS scores and children's academic or social outcomes (Loeb et al. 2004).

Sources: Harms et al. (2005), Harms et al. (2007), Harms et al. (2003), Pianta et al. (2008), Arnett (1989), Clifford et al. (2010), Burchinal et al. (1996), Teachstone (2017), Loeb et al. (2004).

Note: Predictive validity is the extent to which scores on these measures of program quality are predictive of future scores on assessments or measures of children's academic and socio-emotional outcomes.

APPENDIX B

APPROACH TO SYNTHESIZING RESULTS OF VALIDATION STUDIES

Our approach for synthesizing results across the nine states follows the What Works Clearinghouse (WWC) standards version 3.0 (U.S. Department of Education 2014). The WWC released version 4.0 of their standards in October 2017 after we had collected data from author queries and conducted our analyses (U.S. Department of Education 2017). This appendix provides additional details about the key differences between the two versions of the WWC standards, our approach for synthesizing results, and the supplemental analysis that compared the highest and lowest rating levels possible.

Differences between versions 3.0 and 4.0 of the WWC standards

This section describes the key differences between the two versions of the WWC standards. It also discusses how findings might have changed had we used the WWC standards version 4.0.

The approach followed several key steps based on the WWC standards version 3.0. First, we assessed baseline equivalence for each finding related to children's outcomes. Second, we included findings in the analysis if (1) baseline differences were below the WWC cutoff (0.25 standard deviations), and (2) the study used a WWC-approved statistical method to adjust for any baseline differences that fell between 0.05 and 0.25 standard deviations. Under the WWC 3.0 standards, difference-in-difference adjustments and simple gain scores were not approved statistical methods for adjusting for baseline differences, and imputed baseline data could not be used to assess baseline equivalence.

The WWC standards version 4.0 expand the set of approved statistical methods to include difference-in-difference adjustments and simple gain scores and allow for more flexibility in using imputed baseline data, as long as the imputation uses a WWC-approved statistical method (U.S. Department of Education 2017). If we had used the new standards, it is possible that additional findings would have met baseline equivalence standards, and that we might have been able to include more findings related to children's outcomes in our analysis. Given that findings from this report are similar to those in Tout et al. 2017 (which analyzed all findings on children's outcomes, regardless of baseline equivalence), it seems unlikely that including additional findings would have altered this report's conclusions.

Study approach based on the WWC 3.0 standards

This section explains how we assessed how similar children were on the baseline assessments, calculated differences between programs with high and low ratings, determined statistical significance, and combined findings within and across states. This approach was based on the WWC standards version 3.0.

Assessing how similar children are on baseline assessments

The most convincing evidence of a relationship between tiered quality rating and improvement system (TQRIS) ratings and children's outcomes comes from analyses that compare children who had similar skills *before* attending a higher- or lower-rated program (baseline equivalence). To determine which of the analyses conducted by states meet this standard, we calculated the standardized mean difference between groups on *baseline* measures, when available. We calculated the standardized mean difference using Hedges' *g* with an adjustment for small samples. This difference was based on the means of the child development

measure at baseline for higher- and lower-rated programs (y_{H0} and y_{L0}), the respective sample sizes (n_H and n_L), the respective program-level standard deviations at baseline (s_{H0} and s_{L0}), and the small sample size correction $\omega = [1 - s/(4N - 9)]$, with *N* being the total sample size.

$$(1)\ g = \frac{y_{H0} - y_{L0}}{\sqrt{\frac{(n_H - 1)s_{H0}^2 + (n_L - 1)s_{L0}^2}{n_H + n_L - 2}}}$$

Based on standardized mean differences at baseline and the WWC version 3.0 baseline equivalence standards, we classified differences into three categories as follows:

1. Differences meet baseline equivalence if they are between 0.00 and 0.05 standard deviations.
2. Differences meet baseline equivalence if they are greater than 0.05 and less than or equal to 0.25 standard deviations and the analysis included a statistical control for the outcome measured at baseline.
3. Differences do not meet baseline equivalence if they are greater than 0.25 standard deviations.

Calculating differences between programs with high and low ratings

We calculated a standardized difference between higher- and lower-rated programs for all outcome measures. The standardized measure puts all of the outcome measures on the same metric (a standard deviation), enabling us to compare results from analyses that used different measures in the same domain. For example, scores on the Classroom Assessment Scoring System (CLASS) range from 1 to 7 and scores on the Preschool Program Quality Assessment (PQA) range from 1 to 5, so we cannot compare them unless we use a standardized metric. For each outcome measure, we calculated effect sizes using approaches developed for the WWC. Specifically, we used a Hedges' *g* effect size, which calculates the average difference in outcomes between higher- and lower-rated programs based on the analytic method used by the states. The different types of calculations are described below.

Standardized mean difference. For the program quality measures, we calculated the standardized mean difference using Hedges' *g* with an adjustment for small samples. This difference was based on the means of the program quality measure for higher- and lower-rated programs (y_H and y_L), the respective sample sizes (n_H and n_L), the respective program-level standard deviations (s_H and s_L), and the small sample size correction $\omega = [1 - s/(4N - 9)]$, with *N* being the total sample size. The standardized mean difference was defined as

$$(2)\ g = \frac{\omega(y_H - y_L)}{\sqrt{\frac{(n_H - 1)s_H^2 + (n_L - 1)s_L^2}{n_H + n_L - 2}}}$$

Effect size based on regression analysis. For child development outcomes, the estimated mean difference between children in higher- and lower-rated programs was often based on a regression analysis that statistically controlled for the baseline score. If the regression-adjusted means were provided, we calculated the Hedges' *g* using regression-adjusted means on the child development measure ($y_H^{'}$ and $y_L^{'}$) and the unadjusted standard deviations. The effect size using regression-adjusted means was given by

$$(3)\quad g = \frac{\omega\left(y_H^{'} - y_L^{'}\right)}{\sqrt{\frac{\left(n_H - 1\right)s_H^2 + \left(n_L - 1\right)s_L^2}{n_H + n_L - 2}}}$$

Hedges' *g* can also be calculated using information from hierarchical linear models, which were often used in the validation studies to account for the nesting of children within programs. If the level-two coefficient from this model (γ) was provided, the effect size was calculated as

$$(4)\quad g = \frac{\omega\gamma}{\sqrt{\frac{\left(n_H - 1\right)s_H^2 + \left(n_L - 1\right)s_L^2}{n_H + n_L - 2}}}$$

Effect size using a difference-in-differences adjustment. For child development outcomes, if regression-adjusted means or coefficients were not available, we made a difference-in-differences adjustment to Hedges' *g*. Specifically, we computed the numerator as the difference between the baseline and follow-up mean difference for children attending higher-rated programs and the baseline and follow-up mean difference for children attending lower-rated programs. Defining y_{H0} and y_{L0} as the unadjusted baseline means, the difference-in-differences effect size was defined as

$$(5)\quad g = \frac{\omega\left[\left(y_H - y_{H0}\right) - \left(y_L - y_{L0}\right)\right]}{\sqrt{\frac{\left(n_H - 1\right)s_H^2 + \left(n_L - 1\right)s_L^2}{n_H + n_L - 2}}}$$

Gain scores. We approached the use of gain scores in validation study analyses based on the WWC version 3.0 guidance. First, using a gain score in an analysis does not ensure that groups are equivalent at baseline. Such analyses must still demonstrate baseline equivalence using baseline means and standard deviations. Second, using a gain score as the dependent variable in a model does not account for the correlation between scores on the baseline and follow-up assessments. If baseline differences required a statistical adjustment, the analysis had to include the baseline score separately as a covariate. Lastly, effect sizes computed using means and standard deviations of gain scores for children attending higher- and lower-rated programs are not comparable with the effect sizes described earlier because the metric differs. We calculated effect sizes based on gain score means only if the standard deviations of unadjusted follow-up scores were also provided.

Imputation. We approached the use of imputation based on the WWC version 3.0 guidance. First, baseline equivalence cannot be demonstrated using imputed data. Second, researchers may impute missing data for covariates, but not for outcomes. When the analysis used imputed assessment scores, we requested information using only cases with nonmissing assessment scores and used those data for assessing baseline equivalence and calculating effect sizes.

Determining statistical significance

For all differences of program quality and child development outcomes, we calculated the corresponding statistical significance level for each program quality and child outcome difference and corrected for multiple comparisons when states analyzed multiple measures within a domain. We used the Benjamini-Hochberg method to correct for multiple comparisons. This correction lowers the critical *p*-value (from 0.05) for individual comparisons based on the rank order of the *p*-value for that comparison and the total number of comparisons. Because of this correction, some effect sizes with *p*-values that are less than 0.05 might not be significant (because they are compared with a more conservative threshold). See The WWC Procedures and Standards Handbook, version 3.0, for a full description of the correction.

Combining findings

We followed WWC procedures (version 3.0) for combining findings within states. If there was more than one measure in the program quality domain for a state, we calculated the average difference between programs with high and low ratings as the average of the differences. For child development outcomes, we calculated the difference between children in higher- and lower-rated programs in the same way, by taking the average of the differences for the measures that met baseline equivalence standards. For each domain within a state, we calculated the corresponding statistical significance level using the average sample sizes in the higher- and lower-rated groups.

To combine findings across states within a domain, we used a fixed-effect meta-analysis approach. In this approach, the average difference between programs with higher and lower ratings and the corresponding statistical significance level are both estimated by applying a weight to each state equal to the inverse of its within-state variance. We implemented the fixed-effect meta-analysis approach for child development outcomes in which there were findings from two or more states. For each fixed-effect meta-analysis, we also conducted a Q test for heterogeneity in the findings across states, which follows a chi-square distribution.

Supplemental analysis of the highest and lowest rating levels possible

This section presents additional information on the supplemental analysis that compared the highest and lowest rating levels possible. Table B.1 shows the groups of ratings analyzed and Table B.2 presents the findings from this analysis.

Table B.1. Contrasts for supplemental analysis of highest and lowest possible ratings that could be examined in each state

State	Number of rating levels	Intervention condition	Comparison condition	Domains used in average of findings that meet baseline equivalence standard
California	5	Tier 5	Tier 2	Alphabetics General reading achievement Mathematics
Delaware	5	Star 5	Starting with Stars or Star 2	Social-emotional development
Massachusetts	4	Level 3[a]	Level 1	Alphabetics Mathematics Social-emotional development
Ohio	5	5-star	1-star	General reading achievement
Rhode Island	5	5-star	1-star	Social-emotional development
Washington	5	Level 4	Level 2	Alphabetics Mathematics

Source: State validation reports for California, Delaware, Massachusetts, Ohio, Rhode Island, and Washington.

Note: The largest contrast in each state was created using data provided by the authors. Minnesota and Wisconsin did not report disaggregated data by rating level and we did not request that the authors perform additional analyses.

[a] There were only six Level 4 programs in the Massachusetts sample, so the study excluded that level from all statistical analyses.

Table B.2. Findings from the supplemental analysis of highest and lowest possible ratings that could be examined in each state

Domain	Number of states	Meta-analytic effect size for findings that meet baseline equivalence standard (*p*-value)
Alphabetics	3	0.01 (0.88)
General reading achievement	2	-0.12 (0.33)
Mathematics	3	0.09 (0.34)
Social-emotional development	3	0.03 (0.70)

Source: Author calculations.

Note: Positive results for effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

APPENDIX C

DETAILED STATE-SPECIFIC FINDINGS

This appendix contains a detailed description of each state's validation study and its findings. Some of the study information provided was reported by only a subset of states.

California TQRIS validation study

Setting: The study includes center-based programs in the state of California. (The report presents separate supplemental analyses for a small sample of family child care providers.)

Participants: 166 programs, 33 of which are 2-star-rated programs, and 133 of which are 3-, 4-, or 5-star-rated programs; 1,552 children in the sample. The sample included 35 percent of the 472 fully rated programs participating in the California TQRIS at the time of the study.

Program quality outcomes were collected from a subset of classrooms within programs. Child development outcomes were collected from a sample of children in a subset of classrooms within programs.

Sample characteristics: Sample characteristics are not provided for the full sample of 1,552 children. For those included in an analysis of the association between teacher participation in quality improvement activities and child outcomes (1,075 children), females made up 49 percent of the sample and males made up 51 percent of the sample. Nine percent had special needs. Sixty-four percent spoke exclusively Spanish at home or both Spanish and English, whereas fewer than one-third spoke exclusively English at home. Sixty-three percent were assessed as proficient in English.

Authors' statistical approach: Analysis of variance (ANOVA) was used to compare whether average program quality outcomes differed significantly by individual rating level. The authors used hierarchical linear modeling (HLM) regressions of child outcomes on individual rating levels, controlling for baseline scores and child and family characteristics.

The study also examined associations between the program quality measures used in the TQRIS ratings and child outcomes.

Synthesis contrast: The intervention condition was attending a program rated 3, 4, or 5 stars. The comparison condition was attending a program rated 2 stars.

Table C.1. California findings for child development outcomes

Domain	Outcome measure	Sample	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	WJ III Letter-Word Identification	Preschoolers	N=1,524	-0.08 (0.59)	X
General reading achievement	Story and Print Concepts	Preschoolers	N=1,552	-0.18 (0.23)	X
Cognition	Peg Tapping	Preschoolers	N=1,552	0.22 (0.15)	
Mathematics	WJ III Applied Problems	Preschoolers	N=1,499	0.14 (0.36)	X

Source: State validation report for California.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is between 0.05 and 0.25 standard deviations and a statistical adjustment for the baseline score was made. In the case that a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

WJ III = Woodcock Johnson Tests of Achievement.

Table C.2. California findings for program quality outcomes

Domain	Outcome measure	Sample	Sample size	Standardized mean difference (*p*-value)
Program quality	CLASS Classroom Organization	Centers	N = 166	0.27 (0.17)
Program quality	CLASS Emotional Support	Centers	N = 166	2.40* (0.00)
Program quality	CLASS Instructional Support	Centers	N = 166	0.90* (0.00)

Source: State validation report for California.

Note: Positive results for standardized mean difference favor the intervention group; negative results favor the comparison group.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

CLASS = Classroom Assessment Scoring System.

Delaware Stars for Early Success validation study

Setting: The sample includes centers or school-based providers of infant and toddler care, as well as preschool and family child care providers. All providers are located in Delaware.

Participants: 156 programs, 32 of which are Starting with Stars or Star 2-rated programs, and 124 of which are Star 3-, Star 4-, or Star 5-rated programs. 1,012 children in the sample. The study used weights to generalize study findings to the full population of programs participating in Delaware Stars at the time of the study.

Program quality outcomes were collected from a subset of classrooms within programs. Child development outcomes were collected from a sample of children within programs.

Sample characteristics: Children in the sample attended early care programs, were not yet in kindergarten, and were 5 years of age or younger. The authors did not have birthdates for the children but reported the distribution of children in the sample by cohort. Children ranging in age from 2 to 3 years (born September 1, 2011, to August 31, 2012) and eligible to enter kindergarten in fall 2017 accounted for 22 percent of the sample. Children ranging in age from 3 to 4 years (born September 1, 2010, to August 31, 2011) and eligible to enter kindergarten in fall 2016 accounted for 38 percent of the sample. Children ranging in age from 4 to 5 years (born September 1, 2009, to August 31, 2010) and eligible to enter kindergarten in fall 2015 comprised 40 percent of the sample. The child sample was nearly one-half white and non-Hispanic, one-quarter African American, and 15 percent Hispanic. One-third of assessed children were from families earning less than $25,000 per year; the child sample had a proportionately larger share of low-income children relative to the statewide population.

Authors' statistical approach: Comparison of regression-adjusted program quality outcome means separately by rating level. Comparison of child development outcome means that were regression-adjusted for baseline scores and child and family characteristics, separately by rating category.

The study also examined associations between program quality outcomes and child development outcomes for all children and for low-income children. Low-income children were defined as those whose families had incomes of $25,000 or less or those receiving subsidized care.

Synthesis contrast: The intervention condition was attending a program rated Star 3, Star 4, or Star 5. The comparison condition was attending a program rated Starting with Stars or Star 2.

Table C.3. Delaware findings for child development outcomes

Domain	Outcome measure	Sample	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	WJ III Letter-Word Identification	Preschoolers and toddlers	N = 1,012	-0.09 (0.32)	
Comprehension	PPVT	Preschoolers and toddlers	N = 926	-0.06 (0.53)	
Cognition	HTKS	Preschoolers and toddlers	N = 741	0.25* (0.02)	X
Mathematics	WJ III Applied Problems	Preschoolers and toddlers	N = 933	-0.16 (0.08)	
Social-emotional development	DECA Absence of Behavior Problems	Preschoolers and toddlers	N = 776	0.13 (0.22)	X
Social-emotional development	DECA Total Protective Factors	Preschoolers and toddlers	N = 917	-0.11 (0.23)	X

Source: State validation report for Delaware.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

WJ III = Woodcock Johnson Tests of Achievement.

PPVT = Peabody Picture Vocabulary Test.

HTKS = Head-Toes-Knees-Shoulders.

DECA = Devereaux Early Childhood Assessment.

Table C.4. Delaware findings for child development outcomes for low-income children

Domain	Outcome measure	Sample of low-income children	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	WJ III Letter-Word Identification	Preschoolers and toddlers	N = 491	-0.14 (0.25)	X
Comprehension	PPVT	Preschoolers and toddlers	N = 436	-0.16 (0.18)	X
Cognition	HTKS	Preschoolers and toddlers	N = 349	0.09 (0.51)	X
Mathematics	WJ III Applied Problems	Preschoolers and toddlers	N = 417	-0.15 (0.22)	X
Social-emotional development	DECA Absence of Behavior Problems	Preschoolers and toddlers	N = 393	0.11 (0.40)	X
Social-emotional development	DECA Total Protective Factors	Preschoolers and toddlers	N = 439	-0.24 (0.05)	X

Source: State validation report for Delaware.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

WJ III = Woodcock Johnson Tests of Achievement.

PPVT = Peabody Picture Vocabulary Test.

HTKS = Head-Toes-Knees-Shoulders.

DECA = Devereaux Early Childhood Assessment.

Table C.5. Delaware findings for program quality outcomes

Domain	Outcome measure	Sample	Sample size	Standardized mean difference (*p*-value)
Program quality	CIS	Centers and homes	N = 155	0.47* (0.02)
Program quality	CLASS Classroom Organization (pre-K)	Centers and homes	N = 156	0.63* (0.00)
Program quality	CLASS Emotional Support (pre-K)	Centers and homes	N = 156	0.67* (0.00)
Program quality	CLASS Instructional Support (pre-K)	Centers and homes	N = 156	0.31 (0.12)
Program quality	CLASS Emotional and Behavioral Support (toddler)	Centers and homes	N = 108	0.48* (0.03)
Program quality	CLASS Engaged Support for Learning (toddler)	Centers and homes	N=108	0.52* (0.02)
Program quality	PQA	Centers and homes	N=149	0.87* (0.00)

Source: State validation report for Delaware.

Note: Positive results for standardized mean difference favor the intervention group; negative results favor the comparison group. Delaware also calculated group means controlling for a variety of provider and neighborhood-level characteristics. The standardized mean differences and *p*-values using these regression-adjusted means are: CIS: 0.06 (0.75); CLASS Classroom organization (pre-K): 0.24 (0.22); CLASS emotional support (pre-K): 0.02 (0.92); CLASS instructional support (pre-K): -0.19 (0.33); CLASS emotional and behavioral support (toddler): 0.58* (0.01); CLASS engaged support for learning (toddler): 0.39 (0.08); PQA: 0.68* (0.00).

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

CIS = Arnett Caregiver Interaction Scale.

CLASS = Classroom Assessment Scoring System.

PQA = Preschool Program Quality Assessment.

Massachusetts TQRIS validation study

Setting: The study includes center-based programs in Massachusetts exclusively and does not include family-based, school-based, or after-school or out-of-school-time programs.

Participants: 120 programs, 79 of which are Level 1- or Level 2-rated programs, and 41 of which are Level 3-rated programs. 402 children in the sample.

Program quality outcomes were collected from a subset of classrooms within each program. Child development outcomes were collected from a sample of children within each program.

Sample characteristics: Sample characteristics are provided for the total sample of participating children (462), which is larger than the analysis sample. Among the total sample, 22 percent are English language learners, 13 percent are classified as special education, and 55 percent are from families receiving tuition subsidies.

Authors' statistical approach: ANOVA to compare whether average program quality outcomes differed significantly by individual rating level. HLM regressions of children's outcomes on individual rating levels, controlling for baseline scores and child and family characteristics.

Synthesis contrast: The intervention condition was attending a program rated Level 3. The comparison condition was attending a program rated Level 1 or Level 2.

Table C.6. Massachusetts findings for child development outcomes

Domain	Outcome measure	Sample	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	WJ III Letter-Word Identification	Preschoolers	N = 402	-0.05 (0.65)	X
Comprehension	PPVT	Preschoolers	N = 402	0.10 (0.35)	X
Mathematics	WJ III Applied Problems	Preschoolers	N = 402	-0.11 (0.31)	X
Social-emotional development	DECA	Preschoolers	N = 397	0.05 (0.62)	X
Social-emotional development	PLBS	Preschoolers	N = 389	-0.18 (0.10)[b]	

Source: State validation report for Massachusetts.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

[b] Effect size is calculated as a difference-in-differences adjustment to Hedges' g. A statistically adjusted effect size was not calculated because regression-adjusted means or coefficients were not available.

WJ III = Woodcock Johnson Tests of Achievement.

PPVT = Peabody Picture Vocabulary Test.

DECA = Devereaux Early Childhood Assessment.

PLBS = Preschool Learning Behaviors Scale.

Table C.7. Massachusetts findings for program quality outcomes

Domain	Outcome measure	Sample	Sample size	Standardized mean difference (p-value)
Program quality	CIS	Centers (infants and toddlers)	N = 73	0.41 (0.16)
Program quality	CIS	Centers (preschoolers)	N = 120	0.45* (0.02)
Program quality	ECERS-R	Centers	N = 120	0.97* (0.00)
Program quality	ITERS-R	Centers	N = 73	0.57 (0.05)

Source: State validation report for Massachusetts.

Notes: Positive results for standardized mean difference favor the intervention group; negative results favor the comparison group.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

CIS = Arnett Caregiver Interaction Scale.

ECERS-R = Early Childhood Environment Rating Scale-Revised.

ITERS-R = Infant/Toddler Environment Rating Scale-Revised.

Minnesota Parent Aware validation study

Setting: The study includes early care and education programs across Minnesota. Center-based care programs and family child care programs are included.

Participants: 294 programs, 66 of which are 1- or 2-star-rated programs, and 228 of which are 3- or 4-star-rated programs. From 872 to 913 children are in the analysis sample, depending on outcome. The sample included 13 percent of the 2,247 programs participating in Parent Aware at the time of the study.

Program quality outcomes were collected from a subset of classrooms within programs. Child development outcomes were collected from a sample of children in a selected classroom within each program.

Sample characteristics: For the sample of children in participating programs with nonmissing data (568 to 1,181 children depending on the characteristic), the average age in fall was 4.22 years and the average age in spring was 4.65 years. Females made up 49 percent of the sample and males made up 51 percent of the sample. Almost two-thirds (64 percent) were white, 15 percent were African American, 4 percent were Asian, 4 percent were Hispanic, with the remainder in other categories or missing. Children from low-income households made up 62 percent of the sample, 35 percent were from high-income households, and 3 percent were missing data on this variable.

Authors' statistical approach: Comparison of unadjusted program quality means for higher- and lower-rated programs. HLM regression of fall-to-spring gains in child development assessment scores on higher-rated program indicator and child and family characteristics.

The study also examined associations between program quality outcomes and child development outcomes for all children and for low-income children. Low-income children were defined as those from families with incomes at or below 185 percent of the federal poverty level.

Synthesis contrast: The intervention condition was attending a program with a 3- or 4-star rating. The comparison condition was attending a program with a 1- or 2-star rating.

Table C.8. Minnesota findings for child development outcomes

Domain	Outcome measure	Sample	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	TOPEL Phonological Awareness	Preschoolers	N = 872	0.17 (0.07)[b]	
Alphabetics	TOPEL Print Awareness	Preschoolers	N = 898	0.06 (0.49)[b]	
Comprehension	IGDI Picture Naming	Preschoolers	N = 913	0.01 (0.89)[b]	
Cognition	Peg Tapping	Preschoolers	N = 899	0.07 (0.46)[b]	
Mathematics	WJ III Applied Problems	Preschoolers	N = 891	0.02 (0.83)[b]	
Social-emotional development	PLBS Attention-Persistence	Preschoolers	N = 887	0.15 (0.10)[b]	
Social-emotional development	SCBE Anger-Aggression	Preschoolers	N = 908	-0.08 (0.39)[b]	X
Social-emotional development	SCBE Anxiety-Withdrawal	Preschoolers	N = 907	-0.09 (0.34)[b]	
Social-emotional development	SCBE Social Competency	Preschoolers	N = 903	0.29* (0.00)[b]	

Source: State validation report for Minnesota.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

[b] Effect size is calculated as a difference-in-differences adjustment to Hedges' g. A statistically adjusted effect size was not calculated because regression-adjusted means or coefficients were not available.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

TOPEL = Test of Preschool Early Literacy.

IGDI = Individual Growth and Developing Indicators.

WJ III = Woodcock Johnson Tests of Achievement.

PLBS = Preschool Learning Behaviors Scale.

SCBE = Social Competence and Behavior Evaluation.

Table C.9. Minnesota findings for child development outcomes for low-income children

Domain	Outcome measure	Sample of low-income children	Sample size	Statistically adjusted effect size (p-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	TOPEL Phonological Awareness	Preschoolers	N = 512	0.29 (0.10)[b]	
Alphabetics	TOPEL Print Awareness	Preschoolers	N = 533	0.16 (0.33)[b]	
Comprehension	IGDI Picture Naming	Preschoolers	N = 545	0.12 (0.48)[b]	
Cognition	Peg Tapping	Preschoolers	N = 534	0.07 (0.68)[b]	X
Mathematics	WJ III Applied Problems	Preschoolers	N = 526	0.09 (0.61)[b]	
Social-emotional development	PLBS Attention-Persistence	Preschoolers	N = 534	0.05 (0.74)[b]	X
Social-emotional development	SCBE Anger-Aggression	Preschoolers	N = 549	-0.06 (0.72)[b]	
Social-emotional development	SCBE Anxiety-Withdrawal	Preschoolers	N = 548	-0.22 (0.19)[b]	
Social-emotional development	SCBE Social Competency	Preschoolers	N = 546	0.39 (0.02)[b]	

Source: State validation report for Minnesota.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

[b] Effect size is calculated as a difference-in-differences adjustment to Hedges' g. A statistically adjusted effect size was not calculated because regression-adjusted means or coefficients were not available.

TOPEL = Test of Preschool Early Literacy.

IGDI = Individual Growth and Developing Indicators.

WJ III = Woodcock Johnson Tests of Achievement.

PLBS = Preschool Learning Behaviors Scale.

SCBE = Social Competence and Behavior Evaluation.

Table C.10. Minnesota findings for program quality outcomes

Domain	Outcome measure	Sample	Sample size	Standardized mean difference (*p*-value)
Program quality	CLASS Classroom Organization	Centers	N = 261	0.02 (0.91)
Program quality	CLASS Emotional Support	Centers	N = 261	0.07 (0.65)
Program quality	CLASS Instructional Support	Centers	N = 261	0.07 (0.62)
Program quality	ECERS-E Literacy	Centers	N = 145	0.70* (0.00)
Program quality	ECERS-E Literacy	Homes	N = 57	-0.17 (0.53)
Program quality	ECERS-E Mathematics	Centers	N = 145	0.57* (0.00)
Program quality	ECERS-E mathematics	Homes	N = 57	-0.09 (0.75)
Program quality	ECERS-R	Centers	N = 146	0.59* (0.00)
Program quality	FCCERS-R	Homes	N = 57	0.06 (0.82)

Source: State validation report for Minnesota.

Note: Positive results for standardized mean difference favor the intervention group; negative results favor the comparison group.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

CLASS = Classroom Assessment Scoring System.

ECERS-E = Early Childhood Environment Rating Scale-Revised (Extension).

ECERS-R = Early Childhood Environment Rating Scale-Revised.

FCCERS-R = Family Child Care Environment Rating Scale-Revised.

Ohio Step Up To Quality (SUTQ) validation study

Setting: The study includes state-registered early childhood sites in Ohio that had preschool-age enrollment. The sample includes private child care centers, home care providers, and elementary schools.

Participants: 72 programs, 25 of which are 1- or 2-star-rated programs, and 47 of which are 3-, 4-, 5-star-rated programs. From 289 to 325 children in the analysis sample, depending on outcome.

Program quality outcomes were collected from a subset of classrooms within programs. Child development outcomes were collected from all eligible children within programs.

Sample characteristics: Sample characteristics are not provided.

Authors' statistical approach: ANOVA to compare whether average program quality outcomes and average child development outcomes differed significantly by individual rating level and by higher- and lower-rated grouping.

The study also examined associations between program quality outcomes and child development outcomes but did not report any findings.

Synthesis contrast: The intervention condition was attending a program with a 3-, 4-, or 5-star rating. The comparison condition was attending a program with a 1- or 2-star rating.

Table C.11. Ohio findings for child development outcomes

Domain	Outcome measure	Sample	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
General reading achievement	Brigance IED Literacy	Preschoolers	N = 289	-0.03 (0.83)[b]	X
Language development	Brigance IED Language development	Preschoolers	N = 325	0.29* (0.02)[b]	

Source: State validation report for Ohio.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

[b] Effect size is calculated as a difference-in-differences adjustment to Hedges' g. A statistically adjusted effect size was not calculated because regression-adjusted means or coefficients were not available.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

IED = Inventory of Early Development.

Table C.12. Ohio findings for program quality outcomes

Domain	Outcome measure	Sample	Sample size	Standardized mean difference (*p*-value)
Program quality	CHELLO	Homes	N = 16	1.34 (0.02)
Program quality	CIS	Homes	N = 16	0.66 (0.22)
Program quality	CLASS Classroom Organization (pre-K)	Centers	N = 82	0.52 (0.03)
Program quality	CLASS Emotional Support (pre-K)	Centers	N = 82	0.63 (0.01)
Program quality	CLASS Emotional and Behavioral Support (toddler)	Centers	N = 45	0.10 (0.73)
Program quality	CLASS Engaged Support for Learning (toddler)	Centers	N = 45	-0.24 (0.43)
Program quality	CLASS Instructional Support (pre-K)	Centers	N = 82	0.48 (0.05)
Program quality	CLASS Responsive Caregiving (Infant)	Centers	N = 36	0.54 (0.12)
Program quality	ECERS-3	Centers	N = 81	0.44 (0.07)
Program quality	ELLCO General Classroom Environment	Centers	N = 82	0.42 (0.08)
Program quality	ELLCO Language and Literacy	Centers	N = 82	0.42 (0.09)
Program quality	FCCERS-R	Homes	N = 10	1.96 (0.01)
Program quality	ITERS-R	Centers	N = 80	0.60 (0.02)

Source: State validation report for Ohio.

Note: Positive results for standardized mean difference favor the intervention group; negative results favor the comparison group.

CHELLO = Child Home Early Language and Literacy Observation.

CIS = Arnett Caregiver Interaction Scale.

CLASS = Classroom Assessment Scoring System.

ECERS-R = Early Childhood Environment Rating Scale-3.

ELLCO = Early Language and Literacy Classroom Observation.

FCCERS-R = Family Child Care Environment Rating Scale-Revised.

ITERS-R = Infant/Toddler Environment Rating Scale-Revised.

Oregon TQRIS validation study

Setting: The study includes regulated early care and education programs across Oregon, including registered family, certified family, and certified center programs. The programs in the sample served toddlers and preschoolers ages 15 to 60 months.

Participants: 304 programs, 149 of which are Level 1- or Level 2-rated programs, and 155 of which are Level 3-, 4-, or 5-rated programs. The sample included 85 percent of programs fully participating in the Oregon TQRIS rating process at the time of the study.

Sample characteristics: Twenty-one percent of the sample programs were registered family child care programs, 30 percent were certified family, and 49 percent were certified centers.

Program quality outcomes were collected from a subset of classrooms within programs.

Authors' statistical approach: ANOVA to compare whether average program quality outcomes differed significantly by higher and lower rating category.

Synthesis contrast: The intervention condition was attending a program with a Level 3, Level 4, or Level 5 rating. The comparison condition was attending a program with a Level 1 or Level 2 rating.

Table C.13. Oregon findings for program quality outcomes

Domain	Outcome measure	Sample	Sample size	Standardized mean difference (*p*-value)
Program quality	CLASS Classroom Organization	Centers and homes	N = 259	0.42* (0.00)
Program quality	CLASS Emotional Support	Centers and homes	N = 304	0.26* (0.02)
Program quality	CLASS Instructional Support	Centers and homes	N = 304	0.44* (0.00)

Source: State validation report for Oregon.

Note: Positive results for standardized mean difference favor the intervention group; negative results favor the comparison group.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

CLASS = Classroom Assessment Scoring System.

Rhode Island BrightStars validation study

Setting: The study includes center-based early care and education programs that were rated under BrightStars in Rhode Island.

Participants: 71 programs, 42 of which are 1- or 2-star-rated programs, and 29 of which are 3-, 4-, or 5-star-rated programs. 299 to 331 children in the analysis sample, depending on outcome. The sample included 29 percent of the 242 programs participating in BrightStars at the time of the study.

Program quality outcomes were collected from a selected classroom within each program. Child development outcomes were collected from a sample of children within a selected classroom within each program.

Sample characteristics: For the sample of children participating in the study, the average age was 4 years and 4 months, with a range from 36.5 months to 63.7 months. One-quarter of participating families received a subsidy to attend child care. Forty-one percent had an annual household income of $85,000 or more, and more than half of the responding parents from participating families had a bachelor's degree or higher.

Authors' statistical approach: Regression of program quality outcomes on an ordinal star rating variable. HLM regression of child development outcomes on an ordinal star rating variable, controlling for baseline scores and child and family characteristics. HLM regression of child development outcomes on an ordinal star rating and the ordinal star rating interacted with a measure of low-income, controlling for baseline scores and child and family characteristics. Low-income children were defined as those from families with incomes at or below 185 percent of the federal poverty level.

Synthesis contrast: The intervention condition was attending a program with a 3-, 4-, or 5-star rating. The comparison condition was attending a program with a 1- or 2-star rating.

Table C.14. Rhode Island findings for child development outcomes

Domain	Outcome measure	Sample	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	WJ III Letter-Word Identification	Preschoolers	N = 331	0.02 (0.88)[b]	
Alphabetics	WJ III Picture Vocabulary	Preschoolers	N = 331	-0.16 (0.17)[b]	X
Cognition	Pencil Tap Task	Preschoolers	N = 320	0.08 (0.49)[b]	X
Mathematics	WJ III Applied Problems	Preschoolers	N = 327	0.09 (0.41)[b]	X
Social-emotional development	PLBS	Preschoolers	N = 306	-0.11 (0.35)[b]	
Social-emotional development	SCBE Anger-Aggression	Preschoolers	N = 299	0.18 (0.12)[b]	
Social-emotional development	SCBE Anxiety-Withdrawal	Preschoolers	N = 301	0.08 (0.48)[b]	X
Social-emotional development	SCBE Social Competence	Preschoolers	N = 300	0.13 (0.27)[b]	X

Source: State validation report for Rhode Island.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

[b] Effect size is calculated as a difference-in-differences adjustment to Hedges' g. A statistically adjusted effect size was not calculated because regression-adjusted means or coefficients were not available.

WJ III = Woodcock Johnson Tests of Achievement.

PLBS = Preschool Learning Behaviors Scale.

SCBE = Social Competence and Behavior Evaluation.

Table C.15. Rhode Island findings for child development outcomes for low-income children

Domain	Outcome measure	Sample of low-income children	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	WJ III Letter-Word Identification	Preschoolers	N = 100	0.22 (0.28)[b]	X
Alphabetics	WJ III Picture Vocabulary	Preschoolers	N = 100	-0.09 (0.67)[b]	
Cognition	Pencil Tap Task	Preschoolers	N = 98	-0.02 (0.92)[b]	
Mathematics	WJ III Applied Problems	Preschoolers	N = 98	0.10 (0.64)[b]	
Social-emotional development	PLBS	Preschoolers	N = 91	-0.17 (0.43)[b]	
Social-emotional development	SCBE Anger-Aggression	Preschoolers	N = 90	0.23 (0.28)[b]	
Social-emotional development	SCBE Anxiety-Withdrawal	Preschoolers	N = 91	0.23 (0.29)[b]	
Social-emotional development	SCBE Social Competence	Preschoolers	N = 91	0.35 (0.11)[b]	

Source: State validation report for Rhode Island.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

[b] Effect size is calculated as a difference-in-differences adjustment to Hedges' g. A statistically adjusted effect size was not calculated because regression-adjusted means or coefficients were not available.

WJ III = Woodcock Johnson Tests of Achievement.

PLBS = Preschool Learning Behaviors Scale.

SCBE = Social Competence and Behavior Evaluation.

Table C.16. Rhode Island findings for program quality outcomes

Domain	Outcome measure	Sample	Sample size	Standardized mean difference (*p*-value)
Program quality	CLASS Classroom Organization (pre-K)	Centers	N = 71	0.40 (0.10)
Program quality	CLASS Emotional Support (pre-K)	Centers	N = 71	0.53* (0.03)
Program quality	CLASS Emotional and Behavioral Support (toddler)	Centers	N = 32	1.25* (0.00)
Program quality	CLASS Engaged Support for Learning (toddler)	Centers	N = 51	1.43* (0.00)
Program quality	CLASS Instructional Support (pre-K)	Centers	N = 52	0.65* (0.02)

Source: State validation report for Rhode Island.

Note: Positive results for standardized mean difference favor the intervention group; negative results favor the comparison group.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

CLASS = Classroom Assessment Scoring System.

Washington Early Achievers validation study

Setting: The study includes early childhood education programs enrolled in Early Achievers in Washington. The setting included programs for infants, toddlers, and preschoolers in child care centers, family child care programs, Head Start, and Early Childhood Education and Assistance Program sites.

Participants: 100 programs. 100 to 412 children in the analysis sample, depending on outcome. Across the children enrolled in the 100 programs, nineteen percent were enrolled in Level 4-rated programs, 59 percent were enrolled in Level 3-rated programs, and 11 percent were enrolled in Level 2-rated programs. (The remaining children were enrolled in unrated programs and were not included in the synthesis analysis samples.) The sample included 4 percent of the 2,303 programs enrolled in Early Achievers at the time of the study.

Program quality outcomes were collected from a subset of classrooms within programs. Child development outcomes were collected from a sample of children within programs.

Sample characteristics: For the sample of children participating in the study, about half were boys and most spoke English (84 percent). About one-third of the sample were infants or toddlers and two-thirds were preschoolers. About 60 percent were white and 26 percent were another race (14 percent missing). The sample was 14 percent Latino and 72 percent other race (14 percent missing). Twenty-three percent of families reported receiving subsidies. Thirty-five percent of children had parents with at least a bachelor's degree. About one-third of the sample were infants or toddlers and two-thirds were preschool age.

Authors' statistical approach: HLM regression of child development outcomes on rating level, controlling for baseline scores and child and family characteristics.

The study also examined associations between program quality outcomes and child development outcomes.

Synthesis contrast: The intervention condition was attending a program with a Level 3 or Level 4 rating. The comparison condition was attending a program with a Level 2 rating.

Table C.17. Washington findings for child development outcomes

Domain	Outcome measure	Sample	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	EWA Name	Preschoolers	N = 412	-0.06 (0.70)	X
Alphabetics	EWA Words	Preschoolers	N = 405	0.01 (0.97)	X
Alphabetics	WJ III Letter-Word Identification	Preschoolers	N = 409	0.06 (0.68)	X
Comprehension	PPVT	Preschoolers	N = 397	0.35* (0.03)	X
Language development	MSEL Expressive Language	Infants and toddlers	N = 155	0.29 (0.18)	X
Language development	MSEL Receptive Language	Infants and toddlers	N = 174	0.32 (0.10)	X
Cognition	HTKS	Preschoolers	N = 396	0.15 (0.32)	X
Cognition	MSEL Visual Reception	Infants and toddlers	N = 166	0.29 (0.15)	X
Mathematics	TEAM	Preschoolers	N = 403	0.20 (0.20)	X
Science	LENS	Preschoolers	N = 159	0.20 (0.33)	X
Social-emotional development	CBCL	Preschoolers	N = 222	-0.07 (0.74)	
Social-emotional development	CBCL	Toddlers	N = 100	0.02 (0.94)	
Motor skills	MSEL Fine Motor	Infants and toddlers	N = 174	0.49* (0.01)	X
Motor skills	MSEL Gross Motor	Infants and toddlers	N = 138	0.32 (0.14)	X

Source: State validation report for Washington.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

EWA = Early Writing Assessment.

WJ III = Woodcock Johnson Tests of Achievement.

PPVT = Peabody Picture Vocabulary Test.

MSEL = Mullen Scales of Early Learning.

HTKS = Head-Toes-Knees-Shoulders.

TEAM = Tools for Early Assessment in Math.

LENS = Lens on Science.

CBCL = Child Behavior Checklist.

Wisconsin YoungStar validation study

Setting: The study includes both family and group child care providers participating in the YoungStar program in Wisconsin.

Participants: 239 classrooms in 155 programs. Of the 239 classrooms, 108 were rated at the 2 Star level, 102 at the 3 Star level, 7 at the 4 Star level, and 22 at the 5 Star level. 603 to 725 children in the analysis sample, depending on outcome.

Program quality outcomes were collected from a subset of classrooms within programs. Child development outcomes were collected from a sample of children within programs.

Sample characteristics: For the sample of children participating in the study, most children were either white (80 percent) or black (16 percent). About 62 percent of children resided in two-parent households, and slightly fewer than half had a parent with at least a four-year college degree. On average, families reported their incomes of $78,787, and a little more than one-quarter of families received child care subsidies. Nearly all children (98 percent) spoke English at home.

Authors' statistical approach: Comparison of regression-adjusted program quality outcome means between higher- and lower-rated groups and among individual rating levels. HLM regression of child development outcomes on rating-level group (higher contrasted with lower and individual rating levels contrasted with one another), controlling for baseline scores, child and family characteristics, and provider type and region.

The study also examined associations between program quality outcomes and child development outcomes.

Synthesis contrast: The intervention condition was attending a program with a 3-, 4-, or 5-star rating. The comparison condition was attending a program with a 2-star rating.

Table C.18. Wisconsin findings for child development outcomes

Domain	Outcome measure	Sample	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	TOPEL	Preschoolers	N = 639	0.01 (0.85)	X
Alphabetics	WJ III Letter-Word Identification	Preschoolers	N = 725	-0.10 (0.20)	X
Cognition	BSRA	Preschoolers	N = 725	-0.05 (0.53)	X
Cognition	HTKS	Preschoolers	N = 725	-0.05 (0.48)	X
Mathematics	WJ III Applied Problems	Preschoolers	N = 725	-0.08 (0.31)	X
Social-emotional development	PLBS	Preschoolers	N = 604	-0.02 (0.84)[b]	
Social-emotional development	SCBE Anger-Aggression	Preschoolers	N = 603	-0.03 (0.68)[b]	
Social-emotional development	SCBE Anxiety-Withdrawal	Preschoolers	N = 603	0.06 (0.44)[b]	
Social-emotional development	SCBE Social Competence	Preschoolers	N = 603	0.07 (0.40)[b]	

Source: State validation report for Wisconsin.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

[b] Effect size is calculated as a difference-in-differences adjustment to Hedges' g. A statistically adjusted effect size was not calculated because regression-adjusted means or coefficients were not available.

TOPEL = Test of Preschool Early Literacy.

WJ III = Woodcock Johnson Tests of Achievement.

BSRA = Bracken School Readiness Assessment.

HTKS = Head-Toes-Knees-Shoulders.

PLBS = Preschool Learning Behaviors Scale.

SCBE = Social Competence and Behavior Evaluation.

Table C.19. Wisconsin findings for child development outcomes for low-income children

Domain	Outcome measure	Sample of low-income children	Sample size	Statistically adjusted effect size (*p*-value)	Used in domain average of findings that meet baseline equivalence standard[a]
Alphabetics	TOPEL	Preschoolers	N = 119	0.08 (0.69)[b]	
Alphabetics	WJ III Letter-Word Identification	Preschoolers	N = 129	-0.03 (0.87)[b]	
Cognition	BSRA	Preschoolers	N = 129	-0.14 (0.43)[b]	
Cognition	HTKS	Preschoolers	N = 128	-0.11 (0.56)[b]	
Mathematics	WJ III Applied Problems	Preschoolers	N = 129	-0.09 (0.61)[b]	
Social-emotional development	PLBS	Preschoolers	N = 99	-0.03 (0.90)[b]	
Social-emotional development	SCBE Anger-Aggression	Preschoolers	N = 98	0.01 (0.96)[b]	
Social-emotional development	SCBE Anxiety-Withdrawal	Preschoolers	N = 98	-0.07 (0.74)[b]	
Social-emotional development	SCBE Social Competence	Preschoolers	N = 98	-0.38 (0.07)[b]	

Source: State validation report for Wisconsin.

Note: Positive results for mean difference and effect size favor the intervention group; negative results favor the comparison group. The effect size is a standardized measure of the effect of an intervention on student outcomes, representing the change (measured in standard deviations) in an average child's outcome that can be expected if the student receives the intervention.

[a] Findings meet the baseline equivalence standard (1) if the baseline difference between groups is less than 0.05 standard deviations or (2) if the baseline difference between groups is 0.05 to 0.25 standard deviations and a statistical adjustment for the baseline score was made. When a statistically adjusted effect size and a difference-in-differences effect size were calculated, the statistically adjusted effect size will be used in the average.

[b] Effect size is calculated as a difference-in-differences adjustment to Hedges' g. A statistically adjusted effect size was not calculated because regression-adjusted means or coefficients were not available.

TOPEL = Test of Preschool Early Literacy.

WJ III = Woodcock Johnson Tests of Achievement.

BSRA = Bracken School Readiness Assessment.

HTKS = Head-Toes-Knees-Shoulders.

PLBS = Preschool Learning Behaviors Scale.

SCBE = Social Competence and Behavior Evaluation.

Table C.20. Wisconsin findings for program quality outcomes

Domain	Outcome measure	Sample	Sample size	Standardized mean difference (*p*-value)
Program quality	ECERS-R/FCCERS-R	Centers and homes	N = 239	0.62* (0.00)

Source: State validation report for Wisconsin.

Note: Positive results for standardized mean difference favor the intervention group; negative results favor the comparison group. Wisconsin also calculated group means controlling for region. The standardized mean difference (and *p*-value) using the regression-adjusted means is 0.64* (0.00).

* Finding is statistically significant after adjusting for multiple comparisons if necessary.

ECERS-R = Early Childhood Environment Rating Scale-Revised.

FCCERS-R = Family Child Care Environment Rating Scale-Revised.

APPENDIX D

DETAILED FINDINGS ABOUT CHALLENGES

This appendix provides complete findings on all of the challenges that we discussed in the interviews with researchers who conducted the validation studies.

Table D.1. Detailed findings about challenges

Challenge	Number of states categorizing as major challenge	Number of states categorizing as minor challenge	Number of states categorizing as not a challenge
Study design and analysis approach			
Deciding which rating levels to validate	1	4	4
Design limited the interpretation of findings	3	6	0
Sample size and representativeness of sample			
Recruiting programs	4	4	1
Recruiting children	1	4	4
Low response rates from programs	0	5	4
Low response rates from children	2	2	5
Attaining sufficient representation across program types and rating levels	5	3	1
Measures and collecting necessary data			
Selecting measures of program quality	0	5	4
Administering measures of program quality	0	2	7
Analyzing measures of program quality	0	2	7
Selecting measures of child development	2	5	2
Administering measures of child development	0	2	7
Analyzing measures of child development	1	4	4
Obtaining administrative data[a]	2	3	3
Analyzing administrative data[a]	1	5	2
Missing data for programs or children	1	6	2
Limited data on family and child characteristics	2	3	4
Schedule and timing of study			
Conducting study before TQRIS were fully implemented[b]	4	2	0
Collecting data in the allotted study time frame	1	5	3

Sources: Interviews with researchers who conducted the state validation reports in California, Delaware, Massachusetts, Minnesota, Ohio, Oregon, Rhode Island, Washington, and Wisconsin.

[a] Applies to only eight states.

[b] Applies to only six states.

TQRIS = tiered quality rating and improvement system.


DEPARTMENT OF EDUCATION
UNITED STATES OF AMERICA