

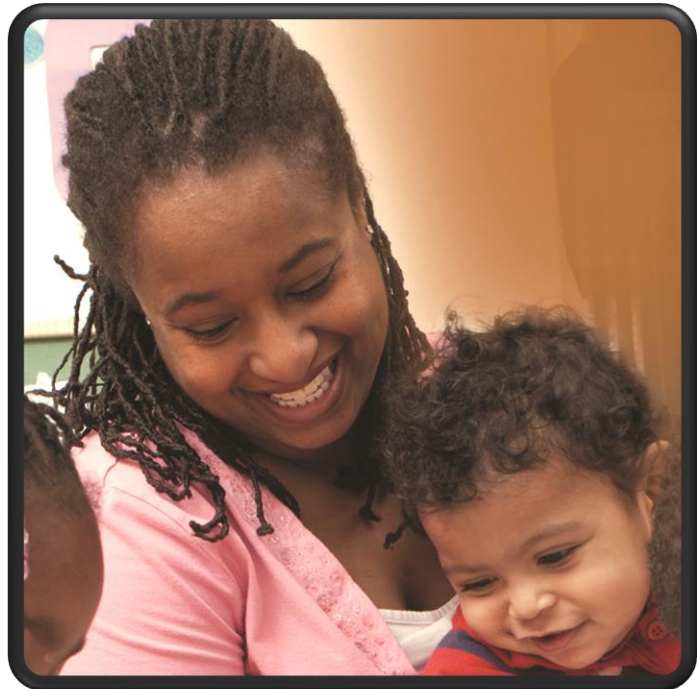
Validation of the QRIS YoungStar Rating Scale

Research Report March 2016

Submitted by Katherine Magnuson, PhD & Ying-Chun Lin, MSW
UW–Madison, School of Social Work and Institute for Research on Poverty

Report 2: Wisconsin Early Child Care Study Findings on the Validity of YoungStar Rating for Children’s School Readiness

The Wisconsin Early Child Care Study (WECCS) is a validation study undertaken to better understand whether Wisconsin’s YoungStar Child Care Quality Rating and Improvement System (QRIS) rating scale is functioning as intended. That is, the study is designed both to explore whether the rating scale is able to differentiate programs according to their levels of observed quality and whether children who attend more highly rated programs gain more in terms of school readiness over the course of a school year than children attending programs rated at lower levels. This report focuses only on the second validity question about whether YoungStar rating predicts differing levels of school readiness among children ages 3 to 5.



BACKGROUND

State policymakers have long been involved in setting the minimum thresholds for structural indicators of child care quality through regulations for licensed providers. As greater attention has been given to the importance of early childhood development as the foundation for later healthy development and learning, states have also increasingly undertaken a range of new initiatives to directly improve children’s early care and education experiences. One type of state policy response has been Tiered Quality Rating and Improvement Systems (TQRIS). In general, TQRIS systems assign early childhood care and education (ECE) providers a rating level, along a quality continuum. This typically serves two important functions. First, it provides a standard way of rating

ECE program quality, based on multiple criteria, and makes the rating information available to parents who will use this information to better select the quality of care they desire. Second, most systems provide a range of technical assistance, resources, and incentives intended to support programs and help them improve their quality. Thus, TQRIS programs attempt to improve the availability and accessibility of high quality ECE by affecting both the demand for high quality care and the supply of such care.

Wisconsin's TQRIS program, YoungStar, was created by a legislative action in the 2009 biennial budget, and it was fully implemented by December of 2010. As with other TQRIS, the key goal is to improve the quality of care that children receive, both by improving parents' knowledge about the quality of specific ECE providers and through supporting providers' efforts to deliver high quality care. The YoungStar rating system assigns providers a star level from 1 to 5 based on measured standardized indicators of quality in four domains: education and professional training, curriculum and learning environment, business and professional practices, and child health and well-being. The Department of Children and Families (DCF) website explains that YoungStar operates in the following way:

- “By objectively measuring child care quality. We rate thousands of child care providers each year, awarding up to five stars for the best quality of care.
- By giving parents an easy way to compare their local child care options and find the programs that match their family's lifestyle, budget, and special needs.
- By supporting providers with tools and training to deliver high-quality care
- By setting a consistent standard for child care quality”

The effectiveness of QRIS programs, in general, is based on the underlying validity of the created rating scales. In this context, validity refers to the ability of rating systems to accurately identify and measure key aspects of quality and program features that may be linked to improvements in children's learning (Zellman & Perlman, 2008). Currently, Wisconsin is one of the 20 states with Race to The Top-Early Learning Challenge (RTT-ELC) federal funding provided specifically to conduct validation studies. Such studies are intended to determine the extent to which there is a “relationship between the ratings generated by the State's Tiered Quality Rating and Improvement System and the learning outcomes of children served by the State's Early Learning and Development Programs” (U.S. Department of Education, 2011). This emphasis on using research to investigate whether the rating scales effectively differentiate program quality and children's learning outcomes is important, as the empirical basis for any one individual quality indicator in predicting classroom quality and children's outcomes is more varied than might be generally appreciated (Burchinal, Magnuson & Powell, 2015). Moreover, careful analyses of the validity of states' overall rating scales with respect to observed classroom quality and children's outcomes are scarce.

In the first report on YoungStar validity (Magnuson & Lin, July 2015), reported results showed that YoungStar program rating levels did predict observed classroom quality

measured by the Environment Rating Scales (ERS), a well-regarded and commonly used measure of classroom quality. In particular, 2 Star programs were rated about 0.5 points lower than 3 to 5 Star programs. The ERS includes a broad set of quality indicators that capture program quality, including the types of materials available, having a regular schedule, the warmth of teacher-child interactions, and health and safety routines. If the ERS measure of observed quality is a strong predictor of children's school readiness, and specifically a strong predictor of children's learning and behavior improvement over time, we might expect that YoungStar ratings would also predict children's school readiness.

However, prior research has provided a mixed view of the extent to which the ERS predicts children's outcomes. An analysis across several large secondary datasets has found very small associations between the overall ERS score and children's outcomes, although somewhat stronger associations between substantively aligned subscales and outcomes (Burchinal, Kainz & Kai, 2011). Moreover, several recent studies of prekindergarten programs have found that the ERS did not predict gains in children's school readiness measures (Mashburn et al., 2008). In general, the explanation for the small to null associations has been that the ERS captures broad domains of child care quality, but not the specific types of interactions, content, or curriculum that are most likely to support children's learning of academic and behavioral skills. In short, scholars have argued that the ERS measure is relevant for measuring general child care quality, but fails to capture the specific practices and content that most strongly lead to better academic and learning-related outcomes.

A handful of validity studies have been done with other states' QRIS programs, and to date only a small number of them includes children's outcomes, although several more are in progress. An early study of Missouri's rating system found that among a small sample of poor preschool children attending center-based programs, QRIS rating predicted higher levels of school readiness, specifically higher receptive vocabulary scores (Thornburg, Mayfield, Hawks, & Fuger, 2009). A more recent study of Keystone STARS, Pennsylvania's QRIS system, found that their two highest rating levels had higher ratings on four-year-old's Work Sampling System score, but direct assessments of children's school readiness were not included in the study (Sirinides, Fantuzzo, LeBoeuf, Barghaus, & Fink, 2015). In contrast, several other studies have found little to no association between QRIS rating level and children's school readiness. In Minnesota, a validity study of Parent Aware found that more highly rated programs differed on just 2 of 8 measures of cognitive and behavioral outcomes (Tout et al., 2016). Similarly, results from a validity study in Colorado did not find consistent associations between QRIS rating levels and preschool-age children's outcomes (Zellman, Perlman, Le, & Setodji, 2008). A study of low-income preschoolers in LA County also found no associations between QRIS ratings and children's cognitive and behavioral outcomes (Soliday Hong, Howes, Marcella, Zucker, & Huang, 2015). Finally, a study by Sabol and colleagues (2013) using prekindergarten classrooms found almost no associations between simulated QRIS rating levels and children's learning outcomes. Taken together, validity studies of other states' QRIS suggest that many of these efforts have not succeeded in strongly differentiating across programs that are most strongly related to children's early learning of academic skills and learning related

behaviors. Nevertheless, the quality indicators and the measurement of these indicators vary widely across states; thus the validity (or lack of validity) of any one state's rating scale does not ensure the validity of other states' rating scales.

This report is the second set of completed analyses from Wisconsin's research validation study of the YoungStar Rating Scale. Starting in 2013, the State of Wisconsin contracted with the Institute for Research on Poverty at UW–Madison to conduct a validation study of the YoungStar rating system. The study was largely funded by the state's RTT-ELC grant and thus shared its emphasis on validity. The overall goal of the study was to examine the validity of the rating scale with respect to both measures of observed classroom quality and children's outcomes. The project was designed with significant input from the DCF. During the course of the study, DCF partners were kept up-to-date on the status of the work and findings, and were consulted on issues as appropriate. The Principal Investigator worked closely with the UW–Madison Survey Center to undertake the data collection.

To preview findings, the results from rigorous analysis of the data indicate that, on average, the children in the validation study are meeting expected benchmarks for learning. The data did not find that children in 3 Star or higher YoungStar programs were demonstrating greater gains than children who were attending 2 Star programs. Children in programs of all levels of rating were demonstrating similar levels of school readiness across multiple early academic and behavioral outcomes in the spring of the study year. Children's school readiness also did not differ when variation in rating points or observational quality measures were used.

SAMPLE AND DATA COLLECTION

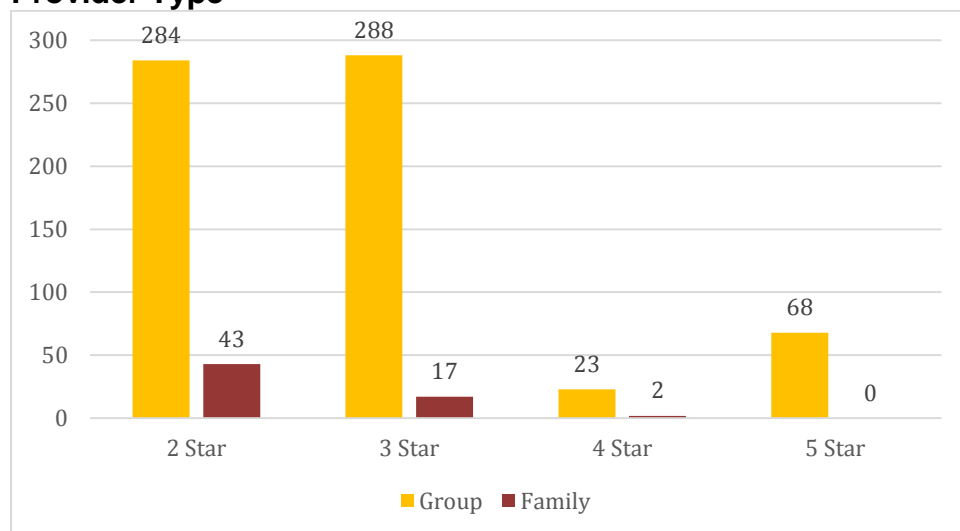
The Wisconsin Early Child Care Study (WECCS). This study was designed to sample both Family and Group child care providers participating in the YoungStar program in May of 2013. The sampling plan was stratified by quality level (low—2 Star and high—3 Star or above) and region (Northeast and Milwaukee) to facilitate comparison across quality levels and ensure representation across types of communities. Within regions, sampling goals were developed to approximate the actual distribution of children ages 3 to 5 across programs by provider type. As a result, a greater number of low-quality family providers were targeted than high-quality family providers (15 2 Star family providers compared with 6 3 Star or higher providers). Programs were asked to participate in the study if they met basic eligibility requirements related to the age of

children served and languages spoken. If the program administrator agreed to participate and at least four children between ages 3 and 5 had completed parental consent forms for their participation, the program was considered enrolled in the study. Figure 1 shows how



children were distributed across YoungStar rating levels, and Appendix 1 provides more details about the recruitment and enrollment of 157 programs and 887 children in the WECCS.

Figure 1: Children Completed Fall and Spring Assessments by Star Rating and Provider Type



A few characteristics of the final study sample are important to keep in mind. First, the study recruited fewer family providers than intended, especially low-quality family providers. Second, the study had lower participation in Milwaukee than in the Northeast region. Nevertheless, the programs enrolled in the study had sufficiently broad coverage to provide a representative sample. Reflecting the distribution of programs in the state, most programs (and classrooms) in this study were in the 2 Star and 3 Star categories.¹

Third, there are multiple pathways to a particular rating. A program may receive an “automated” 2 Star rating, by filling out minimal paperwork with little details about the program and meeting licensing requirements. Twenty-four providers (31%) in our study had this type of automated rating, the remaining 2 Star programs received a “technical” rating based on specific criteria. YoungStar programs receive 4 Star or 5 Star ratings either by meeting the YoungStar criteria during a formal rating process or by automatically receiving this rating if they are accredited by a recognized professional organization (or in the case of Head Start programs by meeting program standards). All of the 5 Star and two of the 4 Star programs in this study had achieved their rating through this type of automated rating.

¹Most programs within the state fall into star ratings 2 and 3. In September of 2013, when data collection began, about 58% of providers were rated at the 2 Star level, 26% at 3 Star, 1% at 4 Star, and about 7% at 5 Star.

In the fall of 2013, 887 parents gave consent for their children who were enrolled in the 157 study programs to participate in child direct assessments. A battery of standardized developmentally appropriate assessments was administered by trained UW Survey Center field staff to collect the baseline information about children's school readiness. Surveys were administered to children's parents, teachers, and program administrators. Children's behaviors were assessed by teacher-report questionnaires during the same data collection period. Once this first wave of data collection was completed, a subset of skilled field workers who had been conducting child assessments were trained to observe classroom quality using the Early Childhood Environment Rating Scale-Revised and the Family Child Care Environment Rating Scale-Revised (ECERS-R and FCCERS-R, see description below, jointly referred to as the Environment Rating Scales, ERS).

In the spring of 2014, another round of direct child assessments was conducted in the same participating child care programs. At this time, 725 children (82%) remained in the study and the same direct child assessments were administered by UW Survey Center field staff in order to evaluate children's gains in academic skills and learning related behaviors. Teachers were also given questionnaires to collect information about their ongoing experiences and children's behavior. Of 725 children having both fall and spring assessments on academic skills in 151 programs, about 90% were rated by teachers on their behaviors (656 in the fall and 644 in the spring).

Table 1 provides descriptive statistics about the children who participated in this study, and for whom direct assessments were conducted in both the fall of 2013 and the spring of 2014. Most children were either White (80%) or Black (16%). About 62% of children resided in two-parent households, and slightly less than half had a parent with at least a 4-year college degree. On average families reported their incomes of \$78,787, and a little more than 1/4 of families received child care subsidies. Nearly all children (98%) spoke English at home. On average children were in these care settings for about 31 hours a week and had been in their current care arrangement for 2 years.

Table 2 provides descriptive statistics about the child care programs sampled in the study and included in this sample for analysis. By design, almost 80% of providers were group providers and 43% of the providers were located in Milwaukee. The average licensed program capacity was 58 children, and the mean classroom size was 15 children. The monthly cost for full-time care was \$748. While a large proportion of providers (40%) had less than one quarter of enrolled children receiving the Wisconsin Shares (state funded child care subsidies for low-income families), 32% of providers had more than 75% of enrolled children receiving the Wisconsin Shares. Among group providers, 34% of lead teachers and 46% of administrators reported that they had at least a bachelor's degree. For lead teachers and administrators with a post-secondary degree, slightly more than a half had special training in early childhood or child development. For group teachers who did not major in Early Childhood Education/Child Development (ECE/CD), the majority of them reported that they had degrees in education or child development related fields of study (e.g., special education). Among family providers, about 15% reported that they had a bachelor's degree or higher; and

again more than half of those with a post-secondary degree had a major in ECE/CD. The mean group size for family providers was 7 children.

MEASURES

Child Developmental Assessments

Child assessments were conducted during one session each in the fall of 2013 and spring of 2014. The assessments were done in a place within the child care setting, such as a hallway or spare conference room, where the child would feel comfortable and safe, but also would be removed from many of the distractions of other children. Testing lasted on average 30 to 40 minutes. For all of the measures, children were given an age appropriate set of first items, and items increased in difficulty. When the child answered several incorrect items in a row, the assessment was completed. UW Survey Center field staff used special software on laptop computers to enter children's responses and minimize administration errors. Spanish versions of the assessments were available for all measures except for the Test of Preschool Early Literacy.

The *Woodcock-Johnson-III Tests of Achievement Letter-Word Identification* (WJLW) was used to measure children's early reading, specifically their letter and word identification skills (Woodcock, Schrank, Mather, & McGrew, 2007). The *Woodcock-Johnson-III Tests of Achievement Applied Problems* (WJAP) was used to assess children's early math skills (Woodcock et al., 2007). The *Bracken School Readiness Assessment-Third Edition* (Bracken) was used to measure children's understanding of basic academic concepts in five categories, including colors, letters, numbers/counting, sizes/ comparison, and shapes (Bracken, 2007). The *Test of Preschool Early Literacy* (TOPEL) subset 3: *Phonological Awareness* was used to assess children's early literacy skills specifically on word elision and blending abilities (Lonigan, Wagner, Torgesen, & Rashotte, 2007). The *Head-Toes-Knees-Shoulders* (HTKS) assessment was used to measure children's behavioral self-regulation. The *Preschool Learning Behaviors Scale* (PLBS) was to assess teacher reports of children's learning behaviors in the child care setting. The teacher version of the *Social Competence and Behavior Evaluation-Short Form* (SCBE-30) was used to assess children's Social Competence, Anger-Aggression, and Anxiety-Withdrawal with 30 items on a 6-point Likert scale (LaFreniere & Dumas, 1995; LaFreniere & Dumas, 1996). Additional details about the measures can be found in Appendix 2.

Other Program, Child, & Family Characteristics

Parents received a phone interview in the fall of 2013. These data are used as the source for information about the child and parent background characteristics. Child and family's characteristics were used in analyses as control variables to minimize the possible confounding effects on program's star rating and children's scores because they have been found to correlate with children's outcomes. These variables included child's gender, race and ethnicity, parent's education level (less than high school, high school/GED, trade school, associate's degree, some college, and bachelor's degree or

Table 1: Children's Characteristics by Star Level

	Total (N = 725)		2 Star (N = 327)		3 Star (N = 305)		4 Star (N = 25)		5 Star (N = 68)	
	% (M)	SD	% (M)	SD	% (M)	SD	% (M)	SD	% (M)	SD
Age (in fall 2013)	4.0	0.6	4.0	0.6	4.0	0.6	3.9	0.6	4.0	0.5
Gender										
Boys	52		48		53		60		59	
Girls	48		52		47		40		41	
Race/Ethnicity										
White	80		80		78		73		91	
Black	16		16		19		20		0	
Hispanic	3		3		2		7		5	
Other/Multi-racial	1		1		1		0		4	
Family Structure										
Two Parents	62		65		63		20		56	
Step Parents	3		5		2		0		5	
Single Parent	31		25		33		73		39	
Others	4		5		3		7		0	
Number of Siblings	1.1	1.0	1.1	1.1	1.2	0.8	0.5	0.6	1.1	0.9
Parent's Education										
Less than High School	3		3		2		0		4	
High School/GED	19		21		18		7		19	
Trade School	6		7		5		0		9	
Associate's Degree	13		17		10		20		11	
Some College	12		10		13		40		11	
Bachelor's Degree or Higher	47		43		52		33		47	
Family's Annual Income	\$78,787	\$63,727	\$74,829	\$52,428	\$83,811	\$72,246	\$47,387	\$49,150	\$83,491	\$70,899
Hours of Care Per Week	30.9	12.6	32.0	11.6	29.4	13.8	35.8	7.6	30.8	12.3
Years in Current Care Provider	2.3	1.4	2.4	1.4	2.3	1.4	2.0	1.4	2.2	1.4
Cost of Current Care (Monthly)	\$410	\$306	\$411	\$288	\$406	\$311	\$322	\$340	\$442	\$355
Receiving Wisconsin Shares	28		26		27		53		35	

Note: Except age and gender, sample size is smaller for other characteristics due to 25% of data missing from parent's survey. Percentage might not add to 100% due to rounding

Table 2: Programs' Characteristics by Star Level

<i>Panel A: Reported at the Program Level</i>	Total (N = 151)		2 Star (N = 77)		3 Star (N = 59)		4 Star (N = 4)		5 Star (N = 11)	
	% (M)	SD	% (M)	SD	% (M)	SD	% (M)	SD	% (M)	SD
In Milwaukee	43		38		49		75		36	
Provider type										
Group providers	79		73		85		75		100	
Family providers	21		27		15		25		0	
Licensed number of children in program	58.4	44.9	47.6	36.5	66.7	51.0	48.0	27.0	93.3	46.0
Monthly full-time care cost for 4-year-old	\$748	\$170	\$700	\$151	\$768	\$177	\$888	\$103	\$931	\$113
Percent of enrolled children receiving Wisconsin Shares										
0-25%	40		40		46		0		30	
26-50%	22		24		15		25		40	
51-75%	5		4		6		25		0	
76% or more	32		31		33		50		30	
<i>Panel B: Reported at the Child Level</i>										
	Total (N = 725)		2 Star (N = 327)		3 Star (N = 305)		4 Star (N = 25)		5 Star (N = 68)	
	% (M)	SD	% (M)	SD	% (M)	SD	% (M)	SD	% (M)	SD
In Milwaukee	39		35		41		68		38	
Provider type										
Group providers	91		87		94		92		100	
Family providers	9		13		6		8		0	
Licensed number of children in program	70.2	44.2	59.0	37.3	76.9	48.5	57.0	15.4	98.5	43.5
Monthly full-time care cost for 4-year-old	\$760	\$167	\$704	\$155	\$775	\$160	\$859	\$80	\$943	\$111
Percent of enrolled children receiving Wisconsin Shares										
0-25%	47		48		53		0		35	
26-50%	21		19		16		32		42	
51-75%	5		4		6		28		0	
76% or more	27		29		24		40		23	

Note: Sample size is smaller for the care cost and percent of children receiving Wisconsin Shares due to 10 – 16% of data missing from administrator's survey. Percentage might not add to 100% due to rounding.

higher), provider's type (group vs. family), provider's region (Milwaukee vs. Northeast) and child's fall score.

YoungStar and Program Quality

Data on YoungStar program quality were provided from YoungStar administrative data. This included the overall rating level of the program as well as the total number of rating points, and the number of points achieved in each of four rating domains: Education and Professional Training (0-15), Learning Environment and Curriculum (0-13), Business and Professional Practices (0-7), and Child Health and Wellbeing (0-5). The average total number of points was 14.88 for this sample (range 2 to 35), and the number of points in each domain was 6.55 for Education and Professional Training, 3.12 for Learning Environment and Curriculum, 4.65 for Business and Professional Practices, and 2.30 for Child Health and Wellbeing. It is important to note that rating points are only available for programs that have had technical or formal rating; those that are rated by an automated pathway do not have information about points available (N = 35).

Of the 157 sites participating in fall data collection, only 2 sites did not participate in the observational component of the study. One of these programs provided wrap around care (and therefore was not open for a sufficient number of hours per day to be observed) and the other had stopped operating. For the remaining 155 programs, the goal was to rate the quality of the classrooms for every child who was assessed in the fall (even if they had moved classrooms between fall and winter). To maximize the number of observations, additional classrooms in participating sites with children ages 3 to 5 were also observed up to a maximum of four classrooms per program. Classrooms serving primarily infants and toddlers were excluded because the observation tool for younger children differs from the one used for older children.

Observed classroom quality was measured by the ERS for center classrooms and family settings. While the rating levels summarize program quality based on a broader range of quality indicators, observed classroom quality focuses on the environment that children experience directly including materials, activities, and interactions with teachers or staff. Specifically, the Early Childhood Environment Rating Scale-Revised (ECERS-R) was used for group providers and the Family Child Care Environment Rating Scale-Revised (FCCERS-R) was used for family providers. The current study used the first six subscales as this is the same practice in the YoungStar rating system. The ECERS-R consists of (1) Space and Furnishings, (2) Personal Care Routines, (3) Language-Reasoning, (4) Activities, (5) Interaction, and (6) Program Structure. The FCCERS-R has similar subscales but has a subscale for Listening and Talking rather than Language-Reasoning.

Scores for each subscale were computed by averaging across the items and the total score was the average score from the six subscale scores. The score ranges from 1 to 7, indicating 1 as inadequate, 3 as minimal, 5 as good, and 7 as excellent. The mean total ERS score for group providers was 4.06 (SD = 0.87) and the mean total ERS score for family providers was 4.11 (SD = 0.99). These descriptive statistics suggest that the bulk of ERS scores were clustered between scores of 3 and 5, and indeed there were very few classrooms that received less than a 3 (N = 26) or more than a 5 (N = 32).

ANALYTIC APPROACH

The primary validation research question addressed by this study report is whether 3- to 5-year-old children in programs rated at higher YoungStar levels gain more in school readiness during one academic program year compared with similar children in programs rated at lower levels. Answering this question is challenging because children are not randomly sorted into programs of differing quality. If this was the case, we could simply compare children in different quality rating levels to learn how the children differ and attribute any differences to their care experiences. However, national data suggest that more affluent children tend to experience higher quality care because their parents are more able to afford the expense of such care. In our data, we see some evidence of this too. Children in 2 Star programs, on average, had parents with lower levels of education and lower family incomes than children in 3 Star programs. It is also worth noting, however, that the small number of children attending 4 Star programs look especially disadvantaged on some family characteristics; these children have the lowest family incomes and levels of parental education, and the highest rates of single parent households.

For this reason, we used analytic models that include measures of relevant background differences to statistically account for any observed or presumed differences across children in differing levels of YoungStar quality. Most important is the use of fall child assessments in predicting the spring outcomes, as these measures provide the best way to ensure that children with similar levels of skills at the outset of the academic year are being compared. In addition, we also include key family background characteristics such as gender, race, and parental education. These measures help to account for differential patterns of school readiness skills that may be due to family background but are not accounted for by children's fall assessment measures.

We estimated two sets of analyses using different operationalization of a high YoungStar rating. The first set of models compare children in 2 Star rated programs with all children in 3 Star and more highly rated programs. In the second model, children in programs with 2 Star rating were compared separately with children in programs with 3 Star, 4 Star, and 5 Star ratings. The first approach is our preferred approach because there are fewer children sampled in the upper star levels, and the comparison is more accurate and precise when the sample sizes are larger.

We also used alternative operationalization of YoungStar quality by considering both the extent to which total rating points and rating points in each of the four YoungStar rating domains are predictive of children's school readiness. The use of points as a key predictor differs from that of star rating, because there is more variability in points across programs, and programs may have higher levels of points in some area even if they did not achieve a higher star rating. As these domains measure different dimensions of program quality, it is possible that some dimensions are more closely related to children's school readiness than others.

Finally, we also briefly consider whether observed classroom quality measured with the ERS and ERS subscales predict children's outcomes. The ERS is a key component of the YoungStar rating scale, and provides another point of comparison for the extent to which child care quality more generally predicts children's learning during the preschool years.

We combined center- and family-based programs together for analyses, after confirming that the pattern of results did not differ across type of program. To handle the dependence among individual child observations in the same child care provider, hierarchical linear modeling was employed with children nested within providers (Raudenbush & Bryk, 2002). Although some classrooms were also nested within center-based programs, this was not the case for all providers, and specifically family-based providers, and with a relatively small number of children for each program, the estimation of a three-level model was not supported.

As noted, a set of controls were used to minimize biases in estimates of the association between program quality and children's school readiness. These measures included the following taken from the parent survey: children's gender, race and ethnicity, parents' education, provider type (group or family provider), and program's region (Milwaukee or Northeast) were included in analyses as covariates. As these variables are likely to relate to children's assessment scores, by holding them constant in analyses, we can better estimate the unique association between program quality and children's outcomes.

Missing Data

By the spring of 2014 some of the children who had participated in fall data collection had left the study (because they were no longer attending the same child care provider). By spring only 725 out of the original 887 children remained in the study, and comprise the sample for this report. Comparing these 725 children with complete child assessment data to those with missing spring assessments, results indicated that children leaving the sample were more likely to live in Milwaukee. Based on the available parent data, children with missing spring assessments were more likely to be Black or Hispanic, and less likely to be from two-parent households. Moreover, these children's families on average had slightly more children at home and much lower income (by about 40% less), and were more likely to move during last year and receive child care or other subsidies. It is important to consider this when interpreting the study results, in that the data are collected for a sample of children who were more advantaged and more stable in terms of child care experiences compared with those who did not remain in the study.

Of 725 children who comprise our sample for this report, 25% of the children's parents did not complete the parent survey and thus have their background characteristic information missing. Our examination of this data found that children of parents who did not complete the parent survey differed slightly by program's region and star level; they were more likely to be from programs in Milwaukee and at the 4 Star level. Because it is important to retain all observations in analysis in order to reduce the selection bias,

missing data were estimated using the statistical procedure of multiple imputation with a chained equations approach. Five data sets were imputed and used for analysis.

RESULTS

Our first question is whether children’s developmental assessments differ across child care programs of differing ratings. If programs rated at a higher YoungStar rating are better able to develop children’s learning skills, we might expect children in higher rated programs to gain more in terms of school readiness assessments and teachers’ rating of learning-related behaviors. We present the mean level of skills and behaviors in the fall of 2013 and the spring of 2014 by YoungStar program quality level (Figures 2 and 3, see also Appendix 2). The general pattern is quite clear. With the exception of the phonological assessment measure, children in all types of programs are on average above the national norms, and also they are improving over the academic year. This suggests that children in YoungStar programs are learning and making age appropriate gains.



However, the data also suggest children who are participating in programs that are rated a 3 Star or higher do not have appreciably better assessment scores in the fall of 2013, nor do they appear to gain more by the spring of 2014, than children in 2 Star programs. This pattern is verified by the regression results presented in Appendix 3, which indicated that there are not significant associations between children’s school readiness assessments in the spring of 2014 and the rating level of child care program they attend. This pattern holds regardless of whether children in 2 Star programs are compared with all other children or whether children in 2 Star programs are compared separately with children in 3 Star, 4 Star, or 5 Star programs. The one exception to this finding is that compared with children in 2 Star programs, children in 5 Star programs have higher ratings on teacher’s reports of Preschool Learning Behavior Scale.

The next set of analyses presented considered whether the YoungStar rating points, rather than star level, predict children’s school readiness. The results indicate that there are no significant associations between rating points and children’s outcomes in either the fall of 2013 or spring of 2014 (see Appendix 3 Tables). The one exception is that in the fall of 2014, programs’ total rating points predicted a higher score on the HTKS assessment. This pattern holds regardless of whether a minimal set of controls were included or a more extensive set. We also found that when the rating points were broken down into the four rating domains, there were not any meaningful associations between these domain specific rating points and our set of children’s school readiness outcomes.

Figure 2: Children’s School Readiness Assessment Scores, by YoungStar Rating Level.

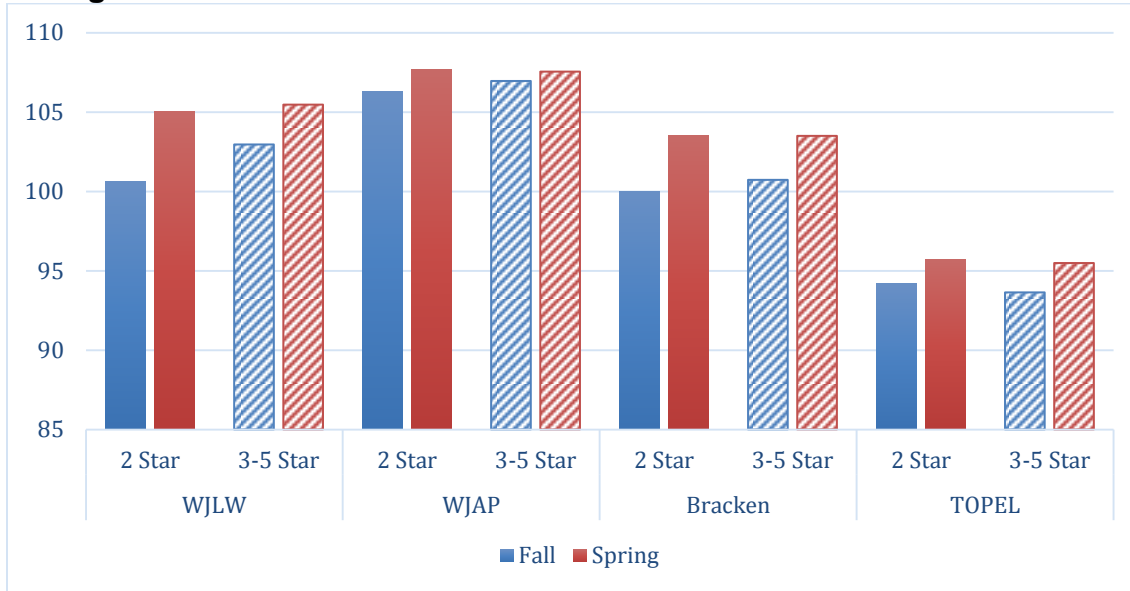
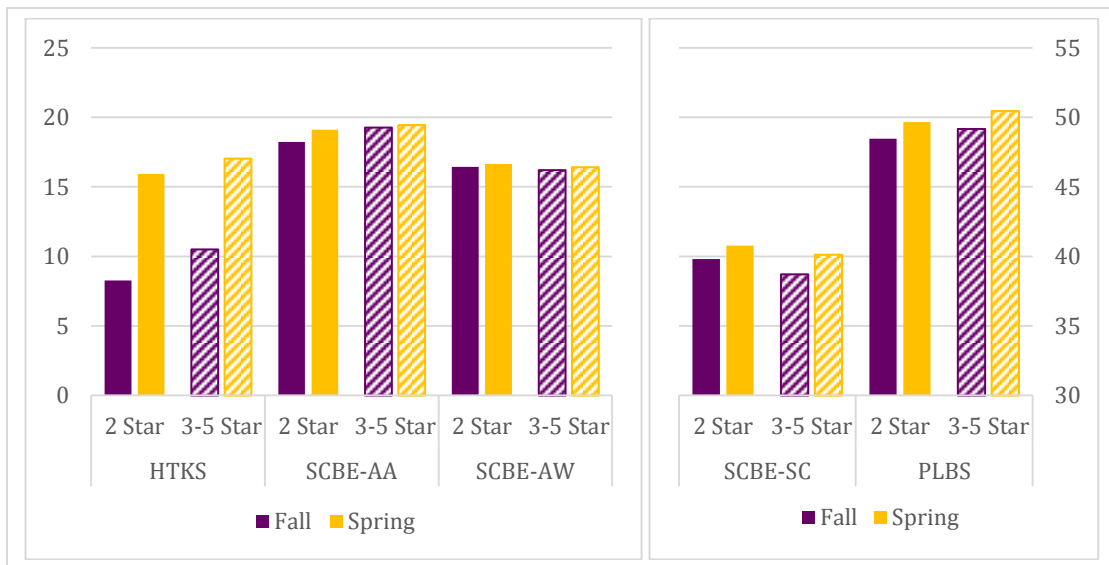


Figure 3: Teacher’s Rating of Children’s Learning Related Behaviors, By YoungStar Rating Level





Our final set of analyses examined whether the observational quality measure used in the YoungStar rating system, the Environment Rating Scales, predicts children's school readiness outcomes. First, we estimated a fairly basic prediction model in which the ERS scores, both the total scores and the individual subscales, predict children's spring assessments taking into account only the region and type of provider. Results are provided in Appendix 3 Tables, and suggest that the ERS rating does predict children's developmental assessments and to a lesser extent

teacher reports of behavior. In particular, three of five skill assessment measures are predicted by the classroom's total ERS score, and all assessments are predicted by the Interaction subscale. On average, higher skills and better classroom behaviors are found among children in classrooms with higher levels of observed quality. And the most predictive dimensions of the classroom are those that are most closely aligned with classroom practices related to learning, a finding supported by the previous literature.

However, results of further analyses indicated that by holding constant a small number of family background characteristics and the child's fall assessment scores, these associations were reduced and no longer statistically significant. Thus, the pattern of associations is likely to reflect not an influence of observed classroom quality, but the fact that more advantaged children experience higher quality classrooms. When children's fall assessments are included as a control variable (Appendix 3 Tables), the results show no association between the ERS measures and children's developmental assessment outcomes or teachers' reports of their behavior. The one exception is that we found that higher rating on the ERS Interaction scale predicts a higher teacher rating for social competence.

SUBGROUP ANALYSES

Although the full sample results did not provide any findings to indicate that YoungStar rating level or points were associated with levels of school readiness, we explored whether the overall sample findings differed across particular subgroups defined by several potentially relevant characteristics. In these analyses we looked at particular groups of children and tested whether YoungStar rating level and the ERS were predictive within these groups.

For those programs that had a technical or formal rating, and thus for whom rating points were available, we also estimated models in which points predicted the outcomes among children attending these programs. All of these models included the same set of

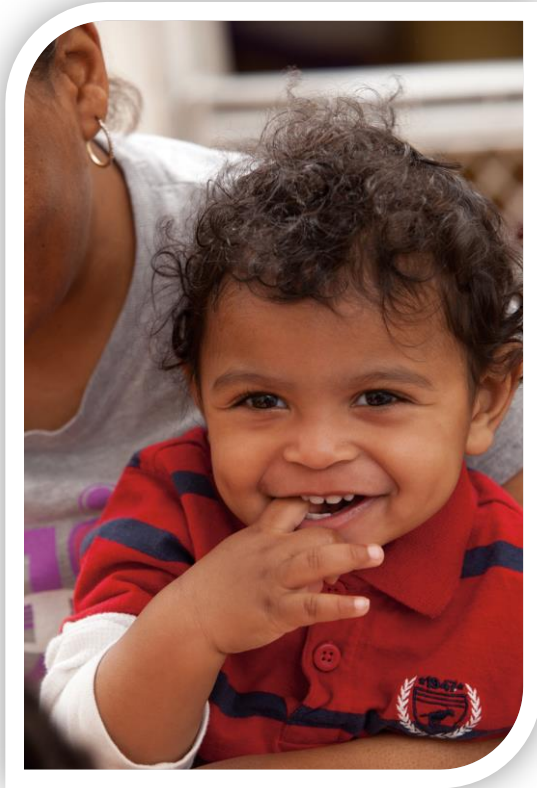
child and family background controls and the child's fall assessment score or teacher's fall report. Because this exercise necessarily reduces the sample size, and thus the power to detect significant effects, any associations to be detected would need to be substantial.

Our first consideration was the level of disadvantage of the children's family background. Prior studies have suggested that program quality might be especially important for disadvantaged children. We looked at subgroups defined by family income, parental education (less than a BA degree) and family structure (single parent), and found no evidence that YoungStar rating level predicted school readiness among children with more disadvantaged backgrounds (or among those with advantaged backgrounds.) We also considered subgroups based on how many hours per week of care the child was in the program, based on the idea that children who had more exposure to the program (20 hour or more of care) may be more influenced by it. We did not find any associations for the subgroup who experienced the care for longer hours (or those who experienced it for shorter hours). We also did not find any associations between YoungStar rating level or points for subgroups defined by group providers or the percentage of children who received child care subsidies. Finally, there were similarly no significant associations between YoungStar rating level or points when the sample was limited to children either in Milwaukee or the Northeast region.

In this type of complex quantitative study, there are numerous specification decisions that are made in order to analyze the data, and sometimes these decisions may affect the pattern of results. Several sensitivity analyses were conducted and compared with the original analyses to evaluate the sensitivity of our reported estimates. The results were robust across these analyses, suggesting that the pattern of results presented was not contingent on a particular coding or modeling strategy (see Appendix 4 for further details).

DISCUSSION

The YoungStar Quality Rating and Improvement System has been operating since 2010. The process of criterion indicator development and implementation was informed by other states' efforts and input from both experts and practitioners. An important goal for Wisconsin has been to use empirical evidence to investigate the extent to which the resulting rating scale and the rating process work as intended to differentiate programs with respect to key indicators of program quality. The WECCS study was undertaken to provide such an examination of the validity of YoungStar rating scale in regards to observed quality and children's school readiness. An earlier report found evidence that YoungStar rating levels predicted observed classroom quality, and this report focuses on findings related to children's school readiness.



WECCS was successful in recruiting and enrolling a sample of community child care providers and children they served in the two selected regions and of varying star ratings and program types. Direct child assessments were completed in the fall of 2013 and spring of 2014. Results from analyses summarized in this report provide answers to questions about the validity of the YoungStar rating scale. First, the data suggest that children enrolled in YoungStar rated programs were performing as well as or better than national norms and developing skills over the period of the study.

The analyses, however, did not find that more highly rated YoungStar programs were generating larger gains in school readiness over the period of the study than programs with a lower rating. This pattern of a lack of association between YoungStar rating and children's outcomes was found across both academic skills and teachers' rating of children's learning related behaviors. These

findings are consistent with most validation studies of other state or local QRIS rating systems that have examined children's outcomes.

In interpreting these findings, it is important to recall that the first report from the validation study found that YoungStar ratings aligned with meaningful differences in observed measures of child care quality, specifically the ECERS-R and FCCERS-R. This suggests that as a measure of child care quality, the YoungStar rating system has achieved validity. The fact that it does not predict differences in specific domains of children's early academic and behavioral skills may suggest that the broad dimensions of child care quality assessed by YoungStar are not the key inputs to further children's pre-academic skills and learning behaviors. This concern has led some states to use other more instructionally attuned measures in their quality rating system, such as the CLASS (Pianta, La Paro & Hamre, 2008). Moreover, greater attention to the implementation of evidence-based curriculum and the details of developmentally appropriate learning activities provided in a developmental sequence may increase the ability of the QRIS system to better differentiate which programs support early academic learning and learning related behaviors.

It is also important to recognize several key limitations in the current study. First, the study was not designed to study the quality of care provided to infants and toddlers. This is an important omission as research has found that care for younger children is often of lower quality. Second, as noted throughout, most children in the study were experiencing 2 Star or 3 Star programs, and thus the primary comparisons were comparing children in these two levels of YoungStar rating. If a larger number of 4 or 5 Star programs were studied, it is possible that the findings might have differed. In addition, many of the children in this study had been attending the same child care arrangement for more than one year, and if much of the differences in learning are set during the toddler years, it may be hard to find such differences in growing during the preschool year. Finally, the study included fewer family providers by design and therefore a larger group of family providers would be needed to draw definitive conclusions about how the rating scale may operate for children in these settings.



These data are among the first to describe the school readiness of preschool age children in YoungStar community based child care programs. First, these data suggest that on average these children are demonstrating developmentally appropriate skill levels, and that these skills are improving over the course of their preschool year. Thus, on average, children in YoungStar programs are meeting important benchmarks that suggest they will enter kindergarten ready to learn. Second, analyses of these data do not find that YoungStar rating level or points predicts children's school readiness. That is, results did not indicate that children in 3 Star or more highly rated programs were performing significantly better, after taking into account child and family's background and prior levels of skills.

REFERENCES

- Bracken, B. A. (2007). *Bracken School Readiness Assessment-Third Edition: Examiner's Manual*. San Antonio, TX: Pearson.
- Burchinal, M., Kainz, K & Cai, Y. (2011). How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from large-scale studies of early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle (ed.), *Quality measurement in early childhood settings* (pp. 11-31). Baltimore, MD: Brooks Publishing.
- Burchinal, M., Magnuson, K., Powell, D., & Hong, S. S. (2015). Early child care and education. In R. M. Lerner (Eds.), *Handbook of child psychology and developmental science, Vol. 4* (pp. 223-267). (7th ed.). New Jersey: Wiley.
- Gresham, F. M., & Elliott, S. N. (2008). *Social Skills Improvement System Rating Scales Manual*. Minneapolis, MN: NCS Pearson, Inc.
- LaFreniere, P. J., & Dumas, J. E. (1995). *Social Competence and Behavior Evaluation-Research Edition Short Form (Teacher Version)*. Los Angeles: Western Psychological Services.
- LaFreniere, P. J., & Dumas, J. E. (1996). *Social Competence and Behavior Evaluation in children ages 3 to 6 year: The short form (SCBE-30)*. *Psychological Assessment*, 8(4), 369-377.
- Lonigan, C. J., Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (2007). *Test of Preschool Early Literacy-Examiner's Manual*. Austin, TX: PRO-ED, Inc.
- Magnuson, K., & Lin, Y.-C. (2015). *Wisconsin Early Child Care Study: Validation of the QRIS Young Star rating scale*. Retrieved from http://dcf.wisconsin.gov/youngstar/pdf/ys-validation-study/weccs_full_report.pdf
- Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., ... & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child development*, 79(3), 732-749.
- McDermott, P. A., Green, L. F., Francis, J. M., & Stott, D. H. (2000). *Preschool Learning Behaviors Scale*. Philadelphia, PA: Edumetric and Clinical Science.
- Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System: Manual Pre-K*. Education Review/Reseñas Educativas.
- Ponitz, C. E. C., McClelland, M. M., Jewkes, A. M., Connor, C. M., Farris, C. L., & Morrison, F. J. (2008). Touch your toes! Developing a direct measure of behavioral regulation in early childhood. *Early Childhood Research Quarterly*, 23(2), 141-158.

- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Sirinides, P., Fantuzzo, J., LeBoeuf, W., Barghaus, K., & Fink, R. (2015). *An inquiry into Pennsylvania's Keystone STARS*. Philadelphia, PA: Consortium for Policy Research in Education.
- Sabol, T. J., Hong, S. L., Pianta, R. C., & Burchinal, M. R. (2013). Can rating pre-K programs predict children's learning? *Science*, 341, 845-846.
- Soliday Hong, S. L., Howes, C., Marcella, J., Zucker, E., & Huang, Y. (2015). Quality Rating and Improvement Systems: Validation of a local implementation in LA County and children's school-readiness. *Early Childhood Research Quarterly*, 30, Part B, 227-240. doi: <http://dx.doi.org/10.1016/j.ecresq.2014.05.001>
- Thornburg, K. R., Mayfield, W. A., Hawks, J. S., & Fuger, K. L. (2009). *The Missouri quality rating system school readiness study*. Columbia, MO: Center for Family Policy & Research. Retrieved from <http://www.elcmdm.org/Knowledge%20Center/reports/MOQRSreport.pdf>
- Tout, K., Cleveland, J., Li, W., Starr, R., Soli, M. & Bultinck, E. (2016). *The Parent Aware Evaluation: Initial Validation Report*. Minneapolis, MN: Child Trends.
- U.S. Department of Education. (2011). *Race to the Top Early Learning Challenge application for initial funding*. Retrieved from <http://www2.ed.gov/programs/racetothetop-earlylearningchallenge/applicant-phase-1.html>
- Woodcock, R. W., Schrank, F. A., Mather, N., & McGrew, K. S. (2007). *Wookcock-Johnsons III Tests of Achievement Form C/Brief Battery*. Rolling Meadows, IL: Riverside Publishing.
- Zellman, G. L., Perlman, M., Le, V., & Setodji, C. M. (2008). *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. Retrieved from http://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_MG650.pdf

APPENDIX 1 SUMMARY OF DATA COLLECTION AND TABLES OF DESCRIPTIVE STATISTICS

Sample phone recruitment goals were set as presented in Appendix Table 1 and starting in June of 2013 recruitment of child care providers for WECCS began. Field work to ensure consent from parents of children enrolled in participating provider programs began in late August of 2013.

Appendix Table 1.1: Summary of the Intended Sample (Showing Stratification by Region, Star Level, and Type of Provider)

Region	2 Star	3 to 5 Star	Total
Milwaukee County	15 Fam/30 Grp (n=210 children)	6 Fam/32 Grp (n=204 children)	21 Fam/62 Grp (n=414 children)
Northeastern Region	15 Fam/30 Grp (n=210 children)	6 Fam/32 Grp (n=204 children)	21 Fam/62 Grp (n=414 children)
Total	30 Fam/60 Grp (n=420 children)	12 Fam/64 Grp (n=408 children)	42 Fam/124 Grp (n=828 children)

Note: Fam = Family Provider, each with two 3- to 5-year-olds; Grp = Group Provider, each with six 3- to 5-year-olds.

Recruitment efforts were time consuming in part because even though many providers appeared to be willing to participate when the study was described on the phone, their circumstances had changed by the time the data collection was fielded. In total, 521 providers in the two regions were called and 246 of these providers agreed to participate in the study (Appendix Table 1.2). Of those who did not agree, some providers directly refused to participate and others were deemed ineligible (most often not serving sufficient numbers of children ages 3–5). Detailed accounting of the reasons that providers were not recruited is provided in Appendix Table 1.3.

Recruitment efforts provided an unadjusted response rate for the recruitment calls of 63.2% ((246/(521-132 ineligibles)=389)). As can be seen by the numbers provided in Table 2, 2 Star providers were less likely to participate than 3 to 5 Star providers (and family 2 Star providers were especially unlikely to participate).

Appendix Table 1.2: Site Recruitment Call Completes and Sampled Sites (Call Completes/Calls sampled)

	2 Star		3–5 Star		Total
	Family	Group	Family	Group	
Milwaukee	26/82	46/125	17/48	40/66	129/321
Northeast	29/69	41/65	9/11	38/55	117/200
Totals	55/151	87/190	26/59	78/121	246/521

Appendix Table 1.3: Final Site Recruitment Disposition

Outcome Description	N	%
Site Recruitment Call Completes	246	47.2%
Refusals	117	22.5%
Eligible, Non-Interview Break-off	5	1.0%
Ineligible, Language Barrier	3	0.6%
No Screener Completed	21	4.0%
Ineligible, Not Enough Kids	47	9.0%
Ineligible, Other Reason	60	11.5%
Quota Filled	22	4.2%
Recruitment Total	521	100%

PROVIDER RECRUITMENT—FALL 2013

Field staff training for interviewers was conducted over a two-week period at the end of August 2013. Data collection efforts began in early September, as soon as sites had sufficient numbers of parental consent forms completed which meant that the minimal rates of child participation to be included in the study had been met (2 children per family site and 4 children for center sites).

Once in the field readying for data collection in the fall, 33 sites were also identified as ineligible and an additional 27 refused to participate (or the insufficient number of parents provided consent forms). Given this lower than expected yield from the number of recruited sites, phone site recruitment continued through the fall (and thus the study incurred higher recruitment costs than anticipated). By the end of recruitment, 166 eligible sites agreed to participate, but not all of these sites were able to be completed. In particular, after replenishing the recruitment efforts for the 2 Star family sites, we still had difficulty getting the sites to complete the data collection (most frequently administrators simply did not return calls or cooperate after they had agreed to participate). Data collection efforts continued until the third week of November in efforts to get as many participating sites and children as possible.

By the end of wave one data collection (fall of 2013), 157 sites completed the first round of child assessment collection. Notably, the sampling and recruitment targets were met in the 3 to 5 Star categories, but perhaps not surprisingly, recruitment in the 2 Star category (especially the family providers) fell below the set targets. Nevertheless, among those that agreed to participate in the study, we had a 76% cooperation rate.

Appendix Table 1.4: Eligible, Non-Refusal Fielded Sites (fielded sites/site goals)

	2 Star		3–5 Star		Total
	Family	Group	Family	Group	
Milwaukee	10/15	30/30	7/6	33/32	80/83
Northeast	16/15	33/30	6/6	34/32	89/83
Totals	26/30	63/60	13/12	67/64	169/166

Appendix Table 1.5: Final Fielded Site Disposition

Outcome Description	N	%
Sites completed	157	66.0%
Admins / informants refused	15	6.3%
Parents refused (and unable to reach minimum at site)	12	5.0%
Eligible – unable to recruit parents / not enough time	2	0.8%
Unknown Eligibility – kids bussed in/unable to reach admin	10	2.5%
Sites determined to be ineligible	33	13.9%
Sites acting as cushion holds – not fielded	9	3.8%
Site Total	238	100%

Appendix Table 1.6: Fall 2013 Final Site Completes

	2 Star		3 to 5 Star		Total
	Family	Group	Family	Group	
Milwaukee	7	24	7	30	68/83
Northeast	16	33	6	34	89/83
Totals	23/30	57/60	13/12	64/64	157/166

In the fall 2013, within the 157 sites 887 children completed school readiness assessments (Appendix Table 1.7) and 725 child assessments were completed in the spring of 2014.

Appendix Table 1.7: Fall 2013 Child Assessment Completes

	2 Star		3-5 Stars		Total
	Family	Group	Family	Group	
Milwaukee	15	152	15	198	380
Northeast	39	208	14	246	507
Totals	54	360	29	444	887

Appendix Table 1.8: Spring 2014 Child Assessment Completes

	2 Star		3-5 Stars		Total
	Family	Group	Family	Group	
Milwaukee	10	104	9	160	283
Northeast	33	180	10	219	442
Totals	43	284	19	379	725

In the fall of 2013 about 90% of program and family administrators completed a self-administered questionnaire (SAQ) and 89% of teachers completed an SAQ and completed individual ratings of children’s classroom behavior. The response rates for the spring 2014 SAQs were somewhat lower.

Appendix Table 1.9: Fall 2013 SAQ Response Rates

	Group	Family	Totals
Administrator	109/121	32/36	141/157
Teacher	181/203	N/A	181/203
Child Supp.	723/804	74/83	797/887
Totals	1013/1128	106/119	1119/1247

Appendix Table 1.10: Spring 2014 SAQ Response Rates

	Group	Family	Totals
Administrator	N/A	17/28	17/28
Teacher	114/195	N/A	114/195
Child Supp.	388/650	34/57	422/707
Totals	502/845	51/85	553/930

Appendix 2: Information on Developmental Measures

Child Developmental Assessments

The *Woodcock-Johnson-III Tests of Achievement Letter-Word Identification* (WJLW) was used to measure children's early reading, specifically their letter and word identification skills (Woodcock, Schrank, Mather, & McGrew, 2007). During the administration of the test, the examiner shows individual letters and words and asks the child to identify and pronounce them correctly. The child is not required to understand the meaning of the words. Raw scores were converted into standard scores with a national normative mean of 100 and standard deviation of 15. The average scores for fall and spring assessments are slightly above national norms at 102.21 (SD = 13.64) and 105.64 (SD = 12.48).

The *Woodcock-Johnson-III Tests of Achievement Applied Problems* (WJAP) was used to assess children's early math skills (Woodcock et al., 2007). The child has to listen to a question, identify the relevant information, and then count or do simple addition and subtraction. Raw scores were converted into standard scores with a national normative mean of 100 and standard deviation of 15. For young 3 year olds with very low raw scores (0 or 1), the normative standard scores were not available, so we used the closest scores from the sample (N = 56 for fall and N = 7 for spring). The average scores for fall and spring assessments are substantially above national norms at 107.00 (SD = 13.07) and 108.10 (SD = 12.02).

The *Bracken School Readiness Assessment-Third Edition* (Bracken) was used to measure children's understanding of basic academic concepts in five categories, including colors, letters, numbers/counting, sizes/comparison, and shapes (Bracken, 2007). The child is asked to point to the pictures that show the concepts the examiner asks the child to identify. Raw scores were converted into standard scores based on national norms (M = 100, SD = 15). The average fall and spring scores are 100.80 (SD = 16.83) and 104.12 (SD = 15.19).

The *Test of Preschool Early Literacy* (TOPEL) subset 3: *Phonological Awareness* was used to assess children's early literacy skills specifically on word elision and blending abilities (Lonigan, Wagner, Torgesen, & Rashotte, 2007). To test the elision ability, the child is asked to say a word and then say what is left after dropping out specific sounds. For instance, the child is supposed to answer "tooth", when the examiner asks what is left after dropping "brush" from "toothbrush". To test the blending ability, the child is asked to listen to separate sounds and combine them to form a word. For example, the child is supposed to answer "baseball" when being asked to combine two sounds "base" and "ball" into one word. Spanish version is not available, so the test cannot be administered to monolingual Spanish-speaking children. Raw scores were converted into standard scores based on national norms (M = 100, SD = 15). The mean scores for fall and spring assessments are below these norms at 93.82 (SD = 15.72) and 95.60 (SD = 16.07).

There was an error with the TOPEL software that occurred during the early test administration in the fall, and was corrected when detected. Instead of skipping out after

3 consecutive wrong answers, the software instructed staff to stop after a total of 3 wrong answers. This resulted in 89 out of 725 children having invalid scores, because they should have been asked more questions after answering three wrong answers (not in consecutive order), but were not. Thus, these 89 fall children's scores are excluded from analyses using the TOPEL as outcome.

The *Head-Toes-Knees-Shoulders* (HTKS) assessment was used to measure children's behavioral self-regulation. The task assesses children's attention (a child must focus on the task and instructions being presented), working memory (a child must remember the rules), and inhibitory control (a child must inhibit a dominant response) (Ponitz et al., 2008). Children are asked to play a game in which they have to do the opposite of what the examiner says. In the first part of the task, the examiner instructs children to touch their head, the children are supposed to touch their toes. If children pass the head/toes part of the task, they complete an advanced trial where shoulders/knees are added. The mean scores for fall and spring assessments are 9.5 (SD = 15.91) and 16.54 (SD = 19.59). National norms for the assessment do not exist, so for analyses we standardized the scores by age in six month increments to have a mean of 50 and standard deviation of 10.

Teacher-Reported Behavioral Assessments

The *Preschool Learning Behaviors Scale* (PLBS) was to assess teacher reports of children's learning behaviors in the child care setting. Consisting of 29 items on a 3-point Likert scale, the instrument captures three important dimensions: Competence Motivation, Attention/Persistence, and Attitude Toward Learning (McDermott, Green, Francis, & Stott, 2000). Raw scores were converted into standard scores based on national norms with a mean of 50 and standard deviation of 10. The average fall and spring scores in the sample are 48.86 (SD = 11.10) and 50.11 (SD = 10.50), respectively.

The teacher version of the *Social Competence and Behavior Evaluation-Short Form* (SCBE-30) was used to assess children's Social Competence, Anger-Aggression, and Anxiety-Withdrawal with 30 items on a 6-point Likert scale (LaFreniere & Dumas, 1995; LaFreniere & Dumas, 1996). Separate raw scores were reported for each subscale. The average fall scores are 39.20 (SD = 9.93), 18.81 (SD = 8.91), and 16.31 (SD = 6.12); the average spring scores are 40.41 (SD = 10.37), 19.30 (SD = 8.81), and 16.52 (SD = 6.21). National norms are not available to standardize these scores, so we created standardized scores based on the sample by age in six month age groups, using the sample mean and standard deviation.

Appendix Table 2.1: Children’s Scores on Developmental Assessments by Star Level

	Fall 2013					Spring 2014				
	N	M	SD	Min	Max	N	M	SD	Min	Max
2 Star	327					327				
WJLW Standard Score		100.63	13.38	50	135		105.03	11.75	68	154
WJAP Standard Score		106.32	12.83	79	138		107.69	11.62	74	140
Bracken Standard Score		99.98	16.34	56	142		103.50	14.86	62	143
TOPEL Standard Score		94.17	15.96	55	134		95.73	15.36	57	136
HTKS		8.27	15.31	0	57		15.93	19.59	0	60
HTKS Standard Score		49.09	9.15	26.28	93.61		49.85	10.71	30.88	108.37
3 Star	305					305				
WJLW Standard Score		102.90	13.57	65	167		105.54	12.95	48	157
WJAP Standard Score		107.28	13.17	72	140		107.82	13.77	20	148
Bracken Standard Score		100.48	17.09	56	138		103.73	16.10	40	139
TOPEL Standard Score		93.67	16.02	59	131		96.23	16.74	55	134
HTKS		11.23	16.96	0	57		17.40	19.71	0	59
HTKS Standard Score		51.12	11.16	37.91	153.24		50.24	9.47	30.88	97.78
4 Star	25					25				
WJLW Standard Score		93.76	12.11	68	115		96.8	13.60	64	120
WJAP Standard Score		102.6	12.37	82	126		102.52	10.56	78	123
Bracken Standard Score		90.56	17.12	59	123		93.2	15.75	66	120
TOPEL Standard Score		91.47	15.36	57	110		86.92	17.20	60	124
HTKS		6.6	13.97	0	48		12.64	19.20	0	59
HTKS Standard Score		48.69	8.56	40.28	71.30		47.84	8.10	34.81	70.58
5 Star	68					68				
WJLW Standard Score		106.59	15.16	74	174		108.38	15.75	78	173
WJAP Standard Score		107.16	13.67	82	132		108.19	11.04	82	129
Bracken Standard Score		105.62	16.32	67	133		106.28	15.90	69	132
TOPEL Standard Score		93.44	13.50	59	123		95.22	15.38	57	124
HTKS		8.71	13.99	0	52		17.03	19.28	0	58
HTKS Standard Score		49.84	7.91	40.28	79.06		50.41	9.03	34.81	79.25

Note: TOPEL was not given to non-English speaking children. Children affected by the instrument error in fall were excluded. Two children did not receive HTKS in spring due to program administrators’ refusals.

Appendix Table 2.2: Teacher’s Report on Children’s Behavioral Assessments by Star Level

	Fall 2013					Spring 2014				
	N	M	SD	Min	Max	N	M	SD	Min	Max
2 Star	290					278				
PLBS Standard Score		48.46	11.40	1.00	64.00		49.65	10.64	20.00	64.00
SCBE Social Competence		4.08	0.91	1.70	6.00		4.12	0.95	1.60	6.00
SCBE Anger-Aggression		1.85	0.80	1.00	5.00		1.93	0.82	1.00	5.50
SCBE Anxiety-Withdrawal		1.70	0.63	1.00	4.60		1.69	0.64	1.00	5.30
3 Star	277					273				
PLBS Standard Score		49.33	11.36	1.00	64.00		50.03	10.92	16.00	64.00
SCBE Social Competence		3.98	1.01	1.43	6.00		4.10	1.07	1.20	6.00
SCBE Anger-Aggression		1.98	1.00	1.00	5.30		1.97	0.91	1.00	5.50
SCBE Anxiety-Withdrawal		1.67	0.63	1.00	4.11		1.71	0.65	1.00	4.90
4 Star	25					25				
PLBS Standard Score		45.12	7.26	28.00	62.00		47.44	6.53	31.00	58.00
SCBE Social Competence		3.44	0.86	1.80	5.33		3.62	0.90	1.67	5.40
SCBE Anger-Aggression		2.23	1.06	1.00	5.13		2.13	0.86	1.00	4.10
SCBE Anxiety-Withdrawal		1.68	0.45	1.00	2.50		1.62	0.45	1.00	2.90
5 Star	64					68				
PLBS Standard Score		50.06	9.53	23.00	64.00		53.28	8.77	25.00	64.00
SCBE Social Competence		4.07	1.06	1.50	5.90		4.10	1.15	1.80	5.90
SCBE Anger-Aggression		1.73	0.81	1.00	5.10		1.80	1.02	1.00	5.70
SCBE Anxiety-Withdrawal		1.53	0.53	1.00	3.50		1.47	0.46	1.00	2.80

Note: Different sample sizes between fall and spring are due to missing teacher’s surveys.

APPENDIX 3 REGRESSION RESULTS

Appendix Table 3.1: Star Level Predicting Children’s Spring Standard Scores on Developmental Assessments, Controlling for Fall Scores and Other Characteristics

	WJLW		WJAP		Bracken		TOPEL		HTKS	
	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
Model 1: Low vs. High Star Level										
<i>Girls</i>	0.06	(0.69)	0.43	(0.79)	1.40	(0.80)	1.69	(0.96)	0.94	(0.66)
<i>Parent’s race/ethnicity</i>										
<i>White (reference)</i>										
<i>Black</i>	-0.66	(1.35)	-3.01	(1.39)*	-1.97	(2.05)	-3.95	(1.69)*	-2.23	(1.15)
<i>Hispanic</i>	-2.95	(2.13)	-6.15	(2.88)*	-4.33	(3.32)	-4.10	(2.97)	-2.68	(2.05)
<i>Others</i>	4.66	(3.97)	-0.06	(3.96)	-0.26	(3.83)	-5.03	(4.53)	-2.40	(3.78)
<i>Parent’s education</i>										
<i>High school/GED (reference)</i>										
<i>Less than high school</i>	-1.95	(2.61)	-3.40	(2.87)	-3.11	(3.37)	-4.28	(3.35)	-0.70	(2.52)
<i>Trade school</i>	-3.54	(1.68)*	-2.20	(2.04)	-2.37	(2.11)	-2.10	(2.50)	-3.07	(1.82)
<i>Associate’s degree</i>	0.27	(1.27)	2.26	(1.54)	0.48	(1.69)	-0.65	(1.84)	-0.20	(1.29)
<i>Some college</i>	0.26	(1.37)	1.01	(1.70)	-0.54	(1.58)	-0.56	(1.84)	-1.54	(1.29)
<i>Bachelor’s degree or higher</i>	-0.05	(1.15)	2.97	(1.13)**	0.60	(1.32)	2.38	(1.48)	1.09	(0.92)
<i>Region: Milwaukee</i>	-0.33	(0.99)	-1.57	(1.01)	-2.84	(1.24)*	-0.44	(1.51)	-1.32	(0.90)
<i>Provider type: Family</i>	-2.13	(1.32)	-3.82	(1.44)**	-4.47	(1.54)**	-1.13	(1.92)	-1.47	(1.26)
<i>Fall standard score</i>	0.64	(0.03)**	0.40	(0.03)**	0.61	(0.03)**	0.57	(0.03)**	0.37	(0.03)**
<i>Low: 2 Star (reference)</i>										
<i>High: 3-5 Star</i>	-1.23	(0.83)	-0.95	(0.83)	-0.73	(0.96)	0.24	(1.24)	-0.53	(0.76)
Model 2: Star Level										
<i>2 Star (reference)</i>										
<i>3 Star</i>	-1.09	(0.87)	-0.76	(0.85)	-0.37	(1.00)	0.78	(1.30)	-0.53	(0.80)
<i>4 Star</i>	-4.65	(2.37)	-3.96	(2.25)	-3.83	(2.72)	-6.79	(3.87)	-1.38	(2.16)
<i>5 Star</i>	-0.71	(1.51)	-0.81	(1.46)	-1.40	(1.73)	-0.19	(2.21)	-0.23	(1.37)

Note: Model 2 included all characteristics and fall scores from model 1 as covariates. Children whose TOPEL scores affected by the instrument error were excluded from analyses *P < .05, **P < .01

Appendix Table 3.2: Star Level Predicting Children’s Spring Standard Scores of Teacher-Report Behaviors, Controlling for Fall Scores and Other Characteristics

	PLBS		SCBE-SC		SCBE-AA		SCBE-AW	
	b	(se)	b	(se)	b	(se)	b	(se)
Model 1: Low vs. High Star Level								
Girls	1.76	(0.65)**	2.37	(0.62)***	0.25	(0.60)	-0.49	(0.64)
Parent’s race/ethnicity								
White (reference)								
Black	-2.47	(1.28)	-2.10	(1.11)	0.43	(1.10)	0.30	(1.26)
Hispanic	-3.69	(1.87)*	-1.62	(1.93)	0.24	(1.58)	0.43	(2.20)
Others	-3.27	(3.09)	-0.42	(2.47)	4.42	(2.45)	4.78	(3.14)
Parent’s education								
High school/GED (reference)								
Less than high school	1.46	(2.77)	-0.81	(2.60)	-1.74	(2.06)	-2.80	(2.44)
Trade school	0.26	(1.54)	0.72	(1.37)	0.07	(1.41)	0.85	(2.11)
Associate’s degree	-0.48	(1.44)	0.19	(1.04)	-0.60	(1.07)	-0.32	(1.31)
Some college	0.54	(1.90)	0.05	(1.33)	0.21	(1.08)	-0.27	(2.24)
Bachelor’s degree or higher	2.30	(1.19)	1.41	(0.85)	-1.68	(0.85)	-1.80	(1.30)
Region: Milwaukee	-1.49	(1.06)	-2.22	(1.01)*	0.82	(0.87)	0.55	(1.05)
Provider type: Family	-1.11	(1.30)	1.40	(1.32)	-0.55	(1.15)	-0.39	(1.39)
Fall score	0.52	(0.03)***	0.49	(0.03)***	0.66	(0.03)***	0.51	(0.03)***
Low: 2 Star (reference)								
High: 3-5 Star	-0.13	(0.83)	0.31	(0.88)	-0.10	(0.75)	0.47	(0.87)
Model 2: Star Level								
2 Star (reference)								
3 Star	-0.77	(0.87)	0.38	(0.93)	0.03	(0.79)	1.13	(0.90)
4 Star	0.27	(2.31)	-0.20	(2.60)	-0.79	(1.96)	-0.83	(2.47)
5 Star	2.86	(1.44)*	0.16	(1.62)	-0.47	(1.27)	-2.52	(1.58)

Note: Analyses included imputed data for missing teacher’s report on children’s behavioral assessments in both fall 2013 and spring 2014 (N = 725). Model 2 included all characteristics and fall scores from model 1 as covariates. SCBE scores were standardized within the sample.

*P < .05, **P < .01, ***P < .001

Appendix Table 3.3: Associations between Children’s Standard Scores on Developmental Assessments and Rating Points

	WJLW b (se)	WJAP b (se)	Bracken b (se)	TOPEL b (se)	HTKS b (se)
Model 1: Fall Scores					
Total Rating Points	-0.02 (0.12)	0.10 (0.11)	0.07 (0.15)	0.08 (0.14)	0.15* (0.07)
Model 2: Spring Scores					
Total rating points	-0.11 (0.10)	-0.01 (0.10)	-0.03 (0.13)	0.09 (0.15)	0.03 (0.08)
Model 3: Spring Scores, controlling for Fall Scores, and child and family characteristics					
Fall score	0.61*** (0.03)	0.39*** (0.04)	0.58*** (0.03)	0.56*** (0.04)	0.36*** (0.04)
Total rating points	-0.12 (0.07)	-0.11 (0.07)	-0.09 (0.09)	-0.01 (0.11)	-0.04 (0.07)

Note: Rating points were from rating period of May 2013 and providers who received automated rating were excluded from analyses (N = 580). Children whose TOPEL scores affected by the instrument error in fall 2013 were also excluded from model 1 and 3 (N = 507). HTKS score was standardized within the sample. Models 1 & 2 controlled for region and provider type; Model 3 controls for all measures.

*P < .05, **P < .01, ***P < .001

Appendix Table 3.4: Summary of Regressions of Children’s Standard Scores of Teacher-Report Behaviors on YoungStar Rating Points

	PLBS b (se)	SCBE-SC b (se)	SCBE-AA B (se)	SCBE-AW b (se)
Model 1: Fall scores				
Total Rating Points	0.15 (0.10)	-0.13 (0.09)	0.10 (0.08)	0.00 (0.09)
Model 2: Spring scores				
Total Rating Points	0.11 (0.10)	-0.02 (0.09)	0.05 (0.09)	0.04 (0.10)
Model 3: Spring Scores, controlling for Fall Scores, and child and family characteristics				
Fall score	0.51*** (0.04)	0.48*** (0.03)	0.64*** (0.03)	0.53*** (0.04)
Total Rating Points	0.00 (0.07)	0.02 (0.08)	0.00 (0.06)	0.05 (0.08)

Note: Analyses included imputed data for missing teacher’s report on children’s behavioral assessments in both fall 2013 and spring 2014. Rating points were from rating period May of 2013 and providers who received automated rating were excluded from analyses (N = 580). SCBE scores were standardized within the sample. Models 1 & 2 controlled for region and provider type, Model 3 controls for all measures. *P < .05, **P < .01, ***P < .001

Appendix Table 3.5: Associations between Children’s Spring Standard Scores on Developmental Assessments and ERS Total Score and Individual Subscale

	WJ LW	WJ AP	Bracken	TOPEL	HTKS
	b (se)	b (se)	b (se)	b (se)	b (se)
Total scores	1.50* (0.69)	0.96 (0.66)	2.53** (0.82)	1.47 (0.91)	1.11* (0.51)
Space & Furnishings	0.65 (0.63)	0.77 (0.60)	1.79* (0.75)	0.28 (0.84)	0.30 (0.46)
Personal Care Routines	0.59 (0.61)	0.15 (0.59)	1.15 (0.74)	-0.04 (0.81)	0.50 (0.45)
Activities	0.70 (0.56)	0.16 (0.54)	1.10 (0.67)	1.17 (0.74)	0.67 (0.41)
Interaction	1.19** (0.38)	0.86* (0.37)	1.71*** (0.45)	1.21* (0.50)	0.86** (0.28)
Program Structure	0.60 (0.38)	0.67 (0.37)	1.06* (0.46)	0.54 (0.51)	0.12 (0.29)
Language-Reasoning ^a	0.82 (0.49)	0.82 (0.47)	1.15 (0.60)	1.51* (0.68)	1.08** (0.38)
Listening-Talking ^b	0.77 (1.45)	-2.35 (1.35)	0.83 (1.69)	-1.27 (1.41)	-0.58 (0.82)

Notes:

^aThis subscale is only for group providers.

^bThis subscale is only for family providers. All models controlled for regions and provider type. Children whose TOPEL scores were affected by the instrument error in fall 2013 were excluded from analyses.

*P < .05, **P < .01, ***P < .001

Appendix Table 3.6: Associations between Children’s Spring Standard Scores of Teacher-Report Behaviors and ERS Score and Individual Subscales

	PLBS	SCBE-SC	SCBE-AA	SCBE-AW
	b (se)	b (se)	b (se)	b (se)
Total scores	1.24* (0.62)	0.30 (0.59)	-0.43 (0.59)	-0.05 (0.59)
Space & Furnishings	0.91 (0.57)	0.51 (0.54)	-0.06 (0.52)	-0.39 (0.53)
Personal Care Routines	0.77 (0.57)	-0.26 (0.54)	-0.58 (0.53)	0.16 (0.58)
Activities	1.05* (0.51)	0.33 (0.48)	-0.42 (0.47)	-0.19 (0.49)
Interaction	0.51 (0.34)	0.26 (0.33)	-0.18 (0.32)	-0.02 (0.33)
Program Structure	0.01 (0.34)	-0.35 (0.33)	0.25 (0.33)	0.16 (0.32)
Language-Reasoning ^a	0.82 (0.46)	0.47 (0.46)	-0.53 (0.45)	0.16 (0.43)
Listening-Talking ^b	-0.05 (0.95)	-0.82 (0.79)	0.88 (0.80)	0.74 (0.94)

Note: Analyses included imputed data for missing teacher’s report on children’s behavioral assessments in both fall 2013 and spring 2014 (N = 725). ^aThis subscale is only for group providers. ^bThis subscale is only for family providers. All models controlled for regions and provider type. *P < .05, **P < .01, ***P < .001

Appendix Table 3.7: ERS Scores Predicting Children's Spring Standard Scores, Controlling for Fall Scores and Other Characteristics

	WJLW		WJAP		Bracken		TOPEL		HTKS	
	b	(se)	b	(se)	b	(se)	b	(se)	b	(se)
Model 1: ERS Total Score										
Fall standard score	0.64	(0.03)***	0.41	(0.03)***	0.61	(0.03)***	0.58	(0.03)***	0.36	(0.03)***
ERS total score	0.02	(0.49)	-0.33	(0.50)	0.34	(0.57)	-0.20	(0.69)	0.30	(0.46)
Model 2: Individual Subscale¹										
Space & Furnishings	-0.10	(0.43)	-0.11	(0.44)	0.24	(0.50)	-0.33	(0.62)	-0.21	(0.40)
Personal Care Routines	0.13	(0.42)	-0.01	(0.43)	0.46	(0.49)	0.02	(0.59)	0.41	(0.39)
Activities	-0.06	(0.39)	-0.11	(0.39)	-0.12	(0.46)	-0.32	(0.56)	0.03	(0.37)
Interaction	0.13	(0.27)	0.01	(0.28)	0.37	(0.32)	0.22	(0.39)	0.41	(0.25)
Program Structure	-0.05	(0.27)	-0.10	(0.28)	0.19	(0.31)	-0.23	(0.38)	-0.16	(0.26)
Language-Reasoning ²	-0.04	(0.35)	0.10	(0.34)	0.17	(0.40)	0.45	(0.50)	0.61	(0.33)
Listening-Talking ³	0.52	(0.82)	-1.79	(1.13)	-0.15	(0.92)	-1.30	(0.96)	-0.91	(0.76)

Note: Children whose TOPEL scores were affected by the instrument error in fall 2013 were excluded from analyses. ¹Individual subscale was included in analyses separately, controlling for all characteristics and fall scores from model1. ²This subscale is only for group provider. ³This subscale is only for family provider. *P < .05, **P < .01, ***P < .001

Appendix Table 3.8: ERS Scores Predicting Children’s Spring Standard Scores of Teacher-Report Behaviors, Controlling for Fall Scores and Other Characteristics

	PLBS		SCBE-SC		SCBE-AA		SCBE-AW	
	b	(se)	b	(se)	b	(se)	b	(se)
Model 1: ERS total score								
Fall score	0.52	(0.03)***	0.49	(0.03)***	0.66	(0.03)***	0.51	(0.04)***
ERS total score	0.62	(0.48)	0.59	(0.48)	0.09	(0.42)	0.46	(0.54)
Model 2: Individual subscale^a								
Space & Furnishings	0.52	(0.44)	0.44	(0.44)	0.05	(0.36)	-0.17	(0.47)
Personal Care Routines	0.64	(0.47)	0.00	(0.44)	-0.35	(0.39)	0.18	(0.57)
Activities	0.50	(0.39)	0.39	(0.39)	-0.00	(0.33)	0.16	(0.42)
Interaction	0.16	(0.25)	0.54	(0.26)*	-0.08	(0.23)	0.34	(0.27)
Program Structure	-0.05	(0.26)	-0.21	(0.26)	0.34	(0.23)	0.38	(0.29)
Language-Reasoning ^b	0.43	(0.34)	0.55	(0.35)	-0.08	(0.31)	0.45	(0.37)
Listening-Talking ^c	0.13	(0.71)	0.12	(0.71)	0.49	(0.62)	1.12	(0.84)

Notes: Analyses included imputed data for missing teacher’s report on children’s behavioral assessments in both fall 2013 and spring 2014 (N = 725). SCBE scores were standardized with the sample.

^aIndividual subscale was included in analyses separately, controlling for all characteristics and fall scores from model 1.

^bThis subscale is only for group provider.

^cThis subscale is only for family provider.

*P < .05, **P < .01, ***P < .001

APPENDIX 4 SUMMARY OF SENSITIVITY ANALYSES

Missing Data

First, we considered whether the use of multiple imputation (MI) techniques to handle missing data may have influenced our findings. When only using the complete case data, the results are similar to those using imputed data, although in some cases the coefficients are somewhat larger. In general, the star rating level does not predict children's scores with one exception of 4 Star rating predicting lower reading, math and phonological skills compared with a 2 star rating. Given that this is an unexpected finding, for a small sample, and that these parents of higher rates (40%) of missing information (e.g., race/ethnicity and education), this divergence in findings suggests that this is a chance finding and not substantively meaningful. As for the ERS total scores and rating points, results are similar to those using imputed data, with one exception that the total ERS score is associated with lower math skills. Again, given the unexpected nature of the finding, we attribute this to a chance finding. Finally, we also conducted sensitivity analyses by excluding outliers of children's scores (i.e., 70 and below; 130 and above) and the results were similar to those reported in the Appendix 3 Tables.

For children's behavioral assessments, again, we considered whether the use of multiple imputation (MI) techniques to handle missing data may have influenced our findings. Results from the complete case data were similar to those using imputed data for models with star rating levels, ERS total scores, and rating points.

Number of Control Variables

To better understand whether children's scores are associated with program quality or children with certain family characteristics or skills self-select into programs, we compared models with and without family demographic variables as well as child's lagged dependent variable (i.e., fall assessment scores and teacher reports). Results were similar to those presented here in this report except for models using the ERS scores to predict children's spring scores, when fall measures were not controlled. The ERS scores (particularly the Interaction subscale) still somewhat predicted children's spring scores. However, when further controlling for child's fall score, the total ERS score or any subscale was not predictive.

We also conducted analyses using a parent report on children's behaviors measured in the fall as an additional control predicting teacher report. During the fall survey, parents rated their children's behavior on the Social Skills Improvement System Rating Scales (SSIS). The SSIS measures 7 domains of social skills functioning and 5 domains of problem behaviors (Gresham & Elliott, 2008). In this study, we used the Parent Form and chose 2 subscales (Engagement and Self-Control) to assess social skills; 4 subscales (Externalizing, Bullying, Hyperactivity/Inattention, and Internalizing) to assess problem behaviors. Parents rate each item on the frequency using a 4-point scale (Never, Seldom, Often, Almost Always), and Raw scores were summed for each subscale. The mean score for each subscale is as follows: Engagement (M = 2.38, SD

= 0.45), Self-Control (M = 1.78, SD = 0.52), Externalizing (M = 0.82, SD = 0.41), Bullying (M = 0.38, SD = 0.43), Hyperactivity/Inattention (M = 1.18, SD = 0.57), and Internalizing (M = 0.46, SD = 0.35). Similar to results were obtained when these measures were included as controls to predict teacher report behavioral scores as outcomes. YoungStar rating levels, total rating points, and the ERS scores did not predict children's behaviors as assessed by teachers in the spring.

Alternative Outcomes

In the analyses described above, we modeled children's spring outcomes as a function of control variables and their fall assessment scores. An alternative approach is to use change scores as outcomes (the difference in assessments between spring and fall). We examined whether higher rated YoungStar programs predicted larger changes in children's school readiness from fall to spring (rather than predicting their spring score with the fall score included as a covariate). Again, child and family's background were controlled for in these analyses and the results were similar to models using spring scores as outcomes and controlling for the fall assessment scores. Additional analyses using raw scores as outcomes and controlling for child's age were also conducted; the results remained the same.

Non-Linear Models for Points and ERS scores

We looked for high levels of points vs. lower levels of points. Programs with total points of 20 or higher were compared with programs of points below 20. For the lower level of points, we used the cutoff of 10 points or less. For the ERS we looked at classrooms rated 5 or above compared to below 5. We also looked at classrooms rated 3 or above and compared to those below 3. Results from these alternative models were similar to results from the original models in which neither rating points nor the ERS scores were predictive of children's scores.