Wisconsin Early Child Care Study

Validation of the QRIS YoungStar Rating Scale

Research Report July 2015
Submitted by Katherine Magnuson, PhD & YingChun Lin, MSW
UW–Madison, School of Social Work and Institute for Research on Poverty

Report 1: Wisconsin Early Child Care Study Findings on the Validity of YoungStar Rating for Observed Classroom Quality

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 first validity question about whether YoungStar rating predicts independently observed classroom quality. A second report to be issued by the end of 2015 will present results related to children's outcomes.

Background

State policymakers have long been involved in setting the minimum thresholds for structural indicators of child care quality through regulations for licensing 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 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 (Zelman & 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.

A handful of validity studies have been done with other states' QRIS programs, and only a small number include independent observations of classroom quality. First, a study of Indiana's Pathways to Quality was conducted and found that Indiana's ratings distinguished levels of observed quality, measured by independent ratings on the Environment Rating Scales (ERS, Lahti, Elicker, Zellman, & Fiene, 2015), across a variety of types of community-based providers, with some evidence that the rating scale was especially effective at predicting observed quality among family providers. A similarly designed study of Maine's rating system also yielded consistent support for their lower levels of quality predicting differences in ERS scores compared with higher levels across program types (Lahti et al., 2015). Finally, examination of data from North Carolina providers also found that the state's two highest rating levels had higher ERS scores than the lower levels (Hestenes et al., 2015). Taken together, validity studies of other state QRIS suggest that these efforts, when carefully implemented, are able to provide meaningful information to parents about the quality of ECE programs. However, the quality indicators and the measurement of the indicators vary widely across states; thus the validity of any one state's rating scale does not ensure the validity of other states' rating scales. For this reason, validation research efforts have proliferated, with varying approaches across states.

This report is the first set of completed analyses from Wisconsin's research validation study of the YoungStar rating plan. 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.

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 or 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. Appendix 1 provides more details about the recruitment and enrollment of 157 programs and 239 classrooms or home care settings in the WECCS.

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.

In the fall of 2013, surveys were administered to children's parents, teachers, and program administrators. In addition, children's school readiness was assessed in a battery of standardized assessments administered by trained research staff. 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 Rating Scale-Revised (ECERS-R and FCCERS-R, see description below, jointly referred to as the Environment Rating Scale, ERS).

The ERS training began with field staff taking an online introductory course provided by the creators of the observational scales. The staff came to UW–Madison for two weeks of training (excluding weekends). The training consisted of a careful overview and discussion of the rating scale content during meetings as well as practice observations

[1]Most programs within the state fall into star ratings 2 to 5. 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.

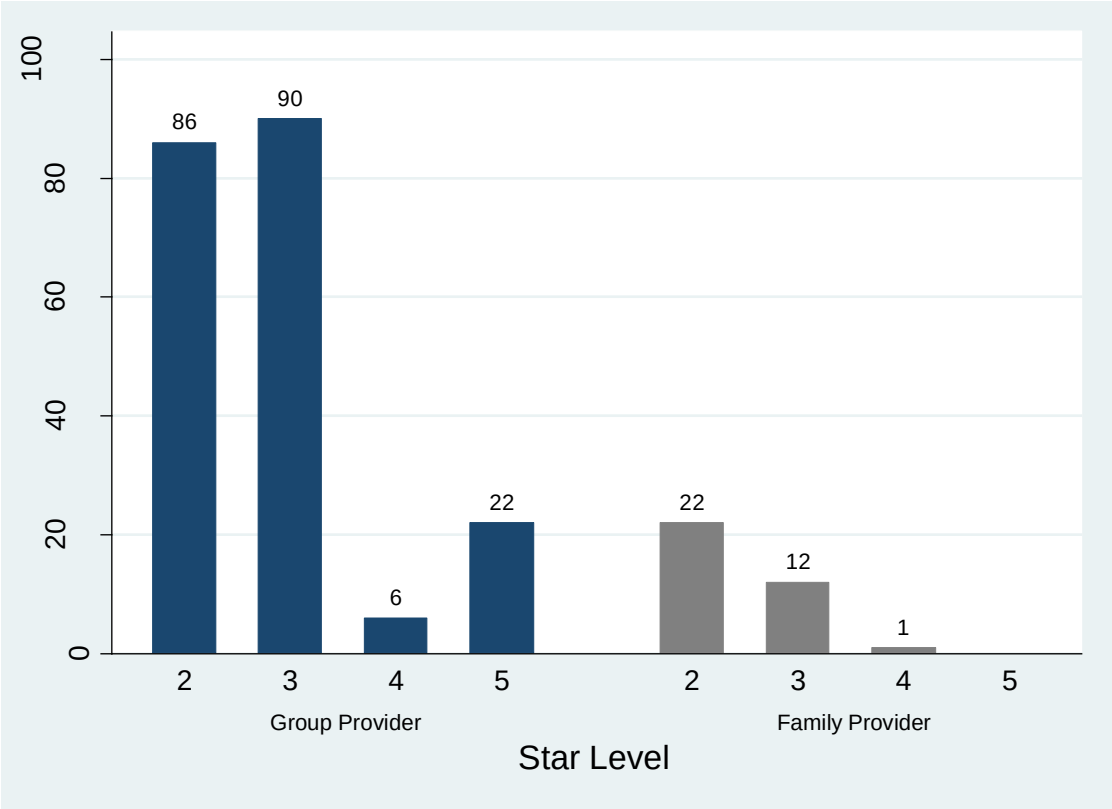
in local Madison child care programs recruited to be of varying YoungStar ratings. To conduct the training, the study brought in master trainers who were experts in the state's YoungStar rating and the goal was to get the raters to be 85% reliable (compared with a master rater) for three consecutive ERS ratings. All six raters who succeeded in training were then employed in the field and their work was reviewed by project staff as it was completed. Raters specialized in only the scale for which they were specifically trained (ECERS-R or FCCERS-R). Observational ratings for the programs began at the very end of December of 2013 and were completed in April of 2014.

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 rating tool for younger children differs from the one used for older children. Of the 155 observed sites (Table 1, Figure 1), there were valid observations of 239 classrooms or family providers. Specifically, 88 (65% of all sites) had only one classroom (or home setting) observed. An additional 67 sites had two classrooms observed, and just 17 had three or four classrooms observed.

Table 1:YoungStar Star Rating for Observed Classrooms, by Region and Provider Type

	Milwaukee County		Northeastern Region	
	N	%	N	%
Family Provider	**13**		**22**	
2 Star	6	46	16	73
3 Star	6	46	6	27
4 Star	1	8	0	0
5 Star	0	0	0	0
Group Provider	**84**		**120**	
2 Star	36	43	50	42
3 Star	37	44	53	44
4 Star	3	4	3	3
5 Star	8	10	14	12
Total	**97**		**142**	

Figure 1: Number of Classrooms Observed by Provider Type and Star Rating



MEASURES

The Early Childhood Environment Rating Scale-Revised (ECERS-R) (Harms, Clifford, & Cryer, 2005) and Family Child Care Environment Rating Scale-Revised (FCCERS-R) (Harms, Cryer, & Clifford, 2007) are observational instruments to measure quality. These rating scales are well validated and have been used frequently in studies of child care quality as well as in the formal rating process for YoungStar.

The ECERS-R is used for center-based or group-based programs and is specific to a classroom and the FCCERS-R is used for family providers in their homes. The ECERS-R consists of seven subscales with total 43 items to measure different dimensions of quality, including (1) Space and Furnishings, (2) Personal Care Routines, (3) Language-Reasoning, (4) Activities, (5) Interaction, (6) Program Structure, and (7) Parents and Staff. The FCCERS-R consists of similar subscales with total 38 items, but differ on the third and the seventh subscales; that is, Listening and Talking and Parents and Providers. The current study uses the subscales 1 to 6 as this is the same practice in the YoungStar rating system.

Classrooms or family providers are observed by a trained rater during at least a 3-hour block and rated based on each indicator in the subscales. The score ranges from 1 to 7, indicating 1 as inadequate, 3 as minimal, 5 as good, and 7 as excellent. Raters give

scores based on the current situation during the observation, not the future plans. If an activity is not observed due to certain condition (e.g. inclement weather and no outdoor activity), teachers or staff are then asked follow-up questions to obtain the information needed. Although all indicators are based in observable phenomena, the YoungStar program in Wisconsin has some specific interpretations to make sure that indicators align with state licensing standards and provide guidance for specific conditions likely to be common in Wisconsin (i.e., snow on playgrounds). For consistency, the field staff were trained to use these interpretations in their ratings for programs.

The six subscale scores are computed by averaging across the items, and the total score is the average score from six subscale sores. The average total ECERS-R score for group providers was 4.08 and the average total FCCERS-R score for family providers was also 4.08, reflecting a level of quality between minimal (ERS=3) and good (ERS=5). The ECERS-R and FCCERS-R average subscales ranged from a low of 2.87 for the Personal Care Routines subscale to 5.01 for the Listening and Talking subscale (for family providers only). Also noteworthy is that some classrooms that were in programs that YoungStar rated as low-quality (2 Star) did achieve high ERS scores when being observed in this study (see Table 2).

There was a minor implementation problem with some of the early observations. In 10 observations, one item in the ECERS-R Interaction subscale related to teachers' supervision during large gross motor activity was recorded by the raters as not being observed. This was most likely due to inclement cold and snowy weather, such that outside play (the most frequent site of gross motor activity) was not observed. In such cases, the rater was supposed to ask relevant questions about how these activities are typically handled, and then provide a rating based on the responses and on the rater's overall impression of the quality classroom supervision (another indicator in the scale). This oversight was corrected in subsequent observations, but the raters did not go back and provide a rating for these ten observations. As a result, the overall ECERS-R scores and Interaction subscales presented in the main analyses for 10 ratings are based on scores in which the gross activity supervision is excluded from the scale.

YoungStar rating level and points were used in analyses as the key explanatory variables (including points in each of the four rating domains). These data were provided by the YoungStar program and include information about the overall YoungStar rating level, as well as points within each rating area for those who had technical or formal ratings (Education and Professional Training, Learning Environment and Curriculum, Business and Professional practices, and Child Health and Wellbeing). For this study, the YoungStar administrative program data including star rating and points within each domain was provided for May of 2013, September of 2013, and May of 2014, representing dates that corresponded with the beginning of recruitment through the end of data collection.

The YoungStar rating scale is set up so that there are both minimum requirements to move from one level to the next, as well as an overall number of points required. There are a total of 40 possible points. At least 11 points are required for a 3 Star rating, 23 points for a 4 Star rating, and 33 points for a 5 Star rating. Two domains account for a

larger number of points. The Education and Training domain accounts for 15 possible points, and the Learning and Curriculum accounts for 13 possible points. Business and Professional Practices have a possible total of 7 points and Child Health and Wellbeing has a possible total of 5 points.

The specific criterion for the rating scale points and minimum requirements differ by program type. Detailed descriptions can be found on the Wisconsin Department of Children and Families website (http://dcf.wisconsin.gov/youngstar/point-detail.htm). Because two of the most important minimum requirements are related to teacher and director education and training as well as ERS observation ratings, these criteria for group providers are highlighted here. The minimum requirement for a 3 Star group provider rating includes having 50% of lead teachers with at least 6 related college credits (verified by The Registry at level 7, a professional development career ladder) and the director must have an administrative credential (Registry level 10). In the Environment and Curriculum domain, a program has to perform self-assessment to achieve a 3 star rating, but there is no requirement or expectation for a minimum ERS rating. For a 4 Star rating, 50% of lead teachers need to have at least 18 related college credits (Registry level 9), and all others have at least 6 credits. In addition, the director must have at least a related AA degree or an unrelated BA degree, and the YoungStar technical staff must complete an ERS rating that results in an average of 4 across classrooms. For a 5 Star rating, all lead teachers should have at least a related AA degree (Registry level 12), and the director must have an administrative credential and either a related AA degree or unrelated BA degree. Finally, for a 5 Star rating, a program must achieve an average of 5 on the YoungStar formal rating ERS observation. In the Business and Professional Practice domain, a 3 Star rating requires ongoing yearly budget, budget review, and accurate tax records. Additional requirements include a written copy of employment policies for a 4 Star rating and evidence of using Model Work Standards for a 5 Star rating. As for the Child Health and Wellbeing, the minimum requirement for all 3 to 5 Star programs is to provide nutritious meals daily.

ANALYTIC APPROACH

The key question examined in this study is the extent to which YoungStar ratings predict observed classroom quality. Better understanding how YoungStar rating levels differentiate observed classroom process quality provides insight into whether the rating scale and its specified criteria are valid.

The distribution of providers across star levels was heavily concentrated in the 2 and 3 Star levels at the time the study was undertaken. Because of this and the fact that primary data collection efforts are costly, the decision was made, in consultation with DCF staff, to focus the study primarily on the question of whether the quality of lower rated programs (2 Star) differed from that of higher rated programs (3 Star or above), and the sampling plan and resulting recruitment reflect this goal (See Appendix 1 for more details about sampling). Table 1 reflects this sampling plan with most classrooms participating in the study being concentrated in the 2 and 3 Star levels, with less representation among 4 and 5 Star rating classrooms. As a result of the sample

distribution, there is sufficient statistical power to detect differences between lower and higher quality programs, but not to detect small or moderate differences among the levels of higher quality programs. For this reason, we focus our analysis and discussion on differences between programs given a 2 Star rating by YoungStar and those given a 3 Star or higher rating.

A secondary question for this study is the extent to which accumulated points in any of the four YoungStar rating domains are especially important in differentiating among classrooms of differing observed quality. As described earlier, four areas in which points are calculated are Education and Professional Training, Learning Environment and Curriculum, Business and Professional Practices, and Child Health and Wellbeing. As these domains measure different types of program characteristics and investments in differing dimensions of program quality, it is possible that some domains may be more predictive of observed classroom quality than others.

To answer the two research questions we primarily rely on conventional regression methods (Ordinary Least Square), in which observed quality (global quality and each subscale) is predicted by measures of a program's star rating. We estimate two models with differing operationalization of the star ratings. In the first estimation model, 2 Star rated programs are compared to all the 3 Star and more highly rated programs. In the second model, the quality of programs at the 2 Star rating is compared separately to programs at the 3 Star, 4 Star, and 5 Star rating. A second set of analyses uses the YoungStar points (overall and in each of the four domains) to predict observed quality. Again, both ERS total score and each subscale are predicted by the measures of YoungStar rating points.

Because programs may have more than one classroom observed, in all regression models we handle the non-independence of these observations by estimating robust standard errors clustered at the program level. In our main analyses, the only control variable included is the measure of the program's region (Milwaukee or Northeast). We do not include other controls, because many of the program and teacher characteristics that might be considered as covariates are likely to contribute to the program's YoungStar rating, and thus should not be held constant.

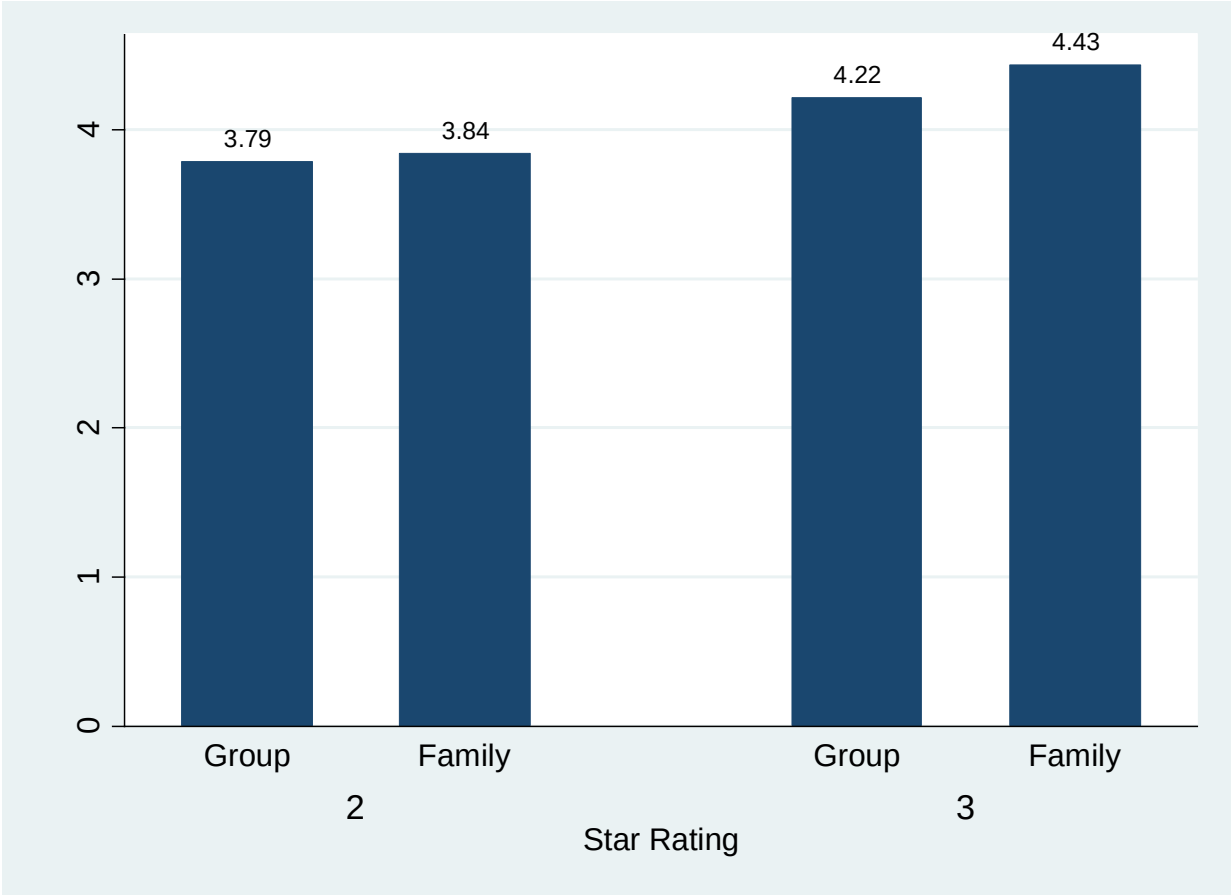
In our analyses, we combined the FCCERS-R and ECERS-R scores, even though they differ slightly in the content of the measured indicators. Our decision was based on our examination of whether the overall ratings differed across program types (group vs. family) within YoungStar rating levels. Table 2 and Figure 2 provide the detailed descriptive statistics, and support the conclusion that the rating scales provide comparable ratings across program types (within star level). As a result, our main analyses in this report present results for all providers, combining family and group providers. Robustness checks conducted separately by program type confirm that the pattern of results provided do not differ substantially across program type.

Table 2: Overall ECERS-R/FCCERS-R Scores by Provider Type and Star Rating

	N	Mean	SD	Min	Max
Group Provider	**204**	**4.08**	**.84**	**1.40**	**5.92**
2 Star	86	3.79	.91	1.40	5.85
3 Star	90	4.22	.72	2.44	5.92
4 Star	6	4.31	.58	3.86	5.26
5 Star	22	4.62	.67	3.24	5.65
Family Provider	**35**	**4.08**	**1.05**	**2.19**	**5.97**
2 Star	22	3.84	1.01	2.19	5.77
3 Star	12	4.43	1.06	2.50	5.97
4 Star	1	5.23	—	—	—
5 Star	0	—	—	—	—

Star ratings were taken from May 2013 data.

Figure 2: Overall ECERS-R/FCCERS-R Scores for 2 and 3 Star Providers by Type



Note: Within star level, the mean ERS scores for group and family providers do not differ.

How Do ERS Scores Differ across YoungStar Rating Levels?

One of the fundamental questions for examining the validity of the YoungStar rating scale is the extent to which a program's rating predicts observed classroom quality. Although, some programs receive their YoungStar rating based on ECERS-R or FCCERS-R observations, many sites do not. In this particular sample, only nine programs' ratings were based on formal ratings using ERS, thus this research is not an exercise in verifying the YoungStar ERS rating process.

The ERS quality observations for programs in this study are significantly higher among more highly rated programs (3 Star or higher) compared with programs rated low quality (2 Star). These differences are apparent when considering mean ERS levels (Figure 3), and when adjustments are made for region in regression analysis (Table 3). Programs with a 2 Star rating have levels of overall observed quality that are above minimal and the ratings of 3 to 5 Star programs are about 0.5 points higher, a rating still below the benchmark of rating of "good." These differences across star levels are highly significant both in the aggregate groupings (2 Star vs. 3 Star or higher), and when comparing each of the 3 Star, 4 Star, and 5 Star groups separately with the 2 Star group.

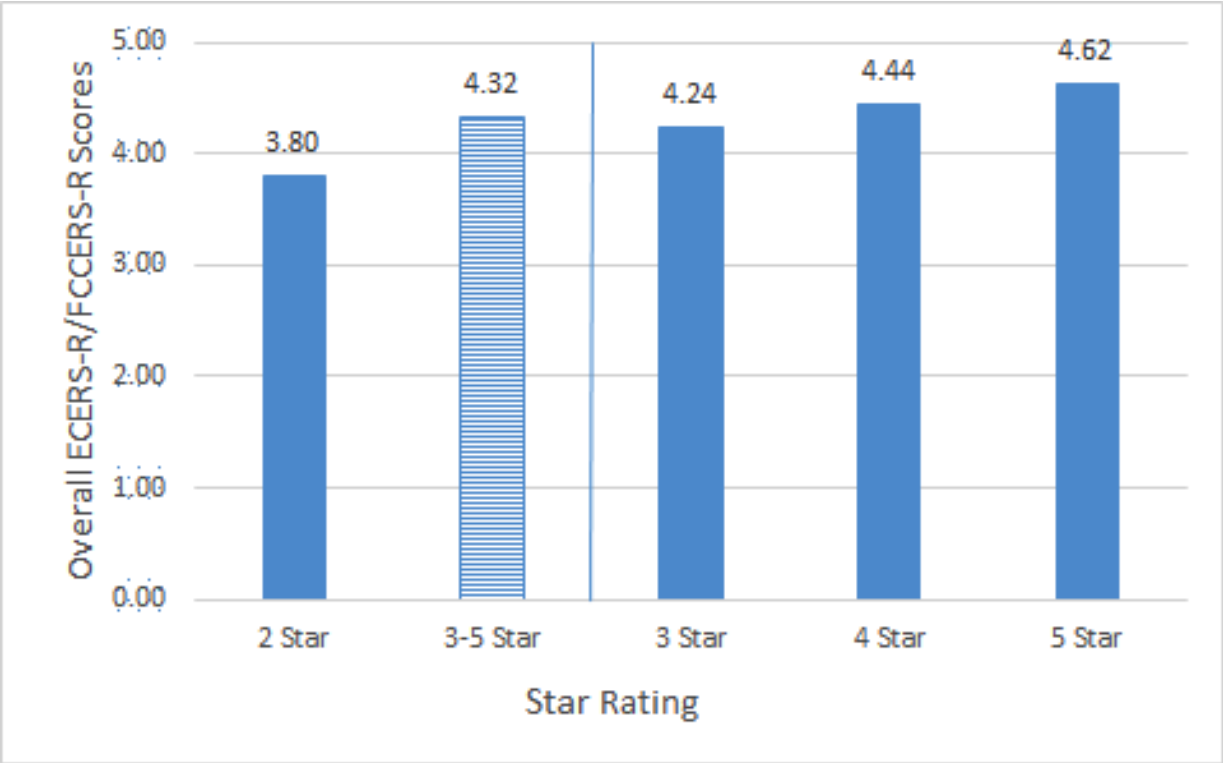
Looking in more details at the higher end of the star rating scale, each star rating is associated with about a 0.2 point higher score scales without adjustment for region (slightly smaller differences when region is taken into account). However, these modest differences are not significantly different from each other (results not shown). This is not surprising given the smaller sample sizes and resulting lower levels of statistical power to detect meaningful differences.

Table 3: Summary of Regressions of Overall Classroom Overall Quality (ECERS-R/FCCERS-R Scores) in YoungStar Rating Level

Star Rating	b	(s.e.)
Model 1: Comparing Low and High Star Ratings		
Low: 2 Star (Reference)	3.91	(.01)
High: 3–5 Star	.53***	(.13)
Model 2: Comparing 4 Star Levels		
2 Star (Reference)	3.91	(.10)
3 Star	.45**	(.13)
4 Star	.70**	(.21)
5 Star	.81***	(.22)

Note: N = 239, *p <.05, **p <.01, ***p < .001. The models controlled for region. The first row under Model 1 and Model 2 shows the regression adjusted average score for the 2 Star classrooms, and the subsequent rows show the regression adjusted difference in score for the relevant higher rated classrooms compared with the 2 Star classrooms. Standard errors of the estimates are in parentheses.

Figure 3: Overall ECERS-R/FCCERS-R Scores by Star Rating



Note: Significant difference in the overall ERS scores comparing high-quality programs (3 Star or above) to low-quality programs (2 Star or lower).

Turning to the subscales of the ERS observations, Figures 4 and 5 illustrate the mean differences in each of the ERS subscales for programs of low and high YoungStar quality rating. Table 4 provides a summary of the regression results. These results suggest that programs rated 3 Star or higher, on average, have higher scores on the following ERS subscales compared with 2 Star programs: space and furnishings, activities, and program structure. In contrast, higher quality programs are not different than lower quality programs in their scores on the ERS interaction and personal care routines subscales. In the case of personal care routines subscale, all programs score relatively low, and in the case of interaction subscale all programs score relatively high.

For group programs, the ECERS-R includes Language and Reasoning subscale and this subscale was significantly higher for classrooms in higher-rated programs compared with 2 Star programs. For family providers, the FCCERS-R includes a Listening and Talking subscale, and the difference between lower and higher quality programs was substantively large among family providers, but with fewer observations there was not sufficient statistical power for it to reach statistical significance.

Figure 4: Summary of ECERS-R/FCCERS-R Subscale Scores for Low and High YoungStar Rated Programs

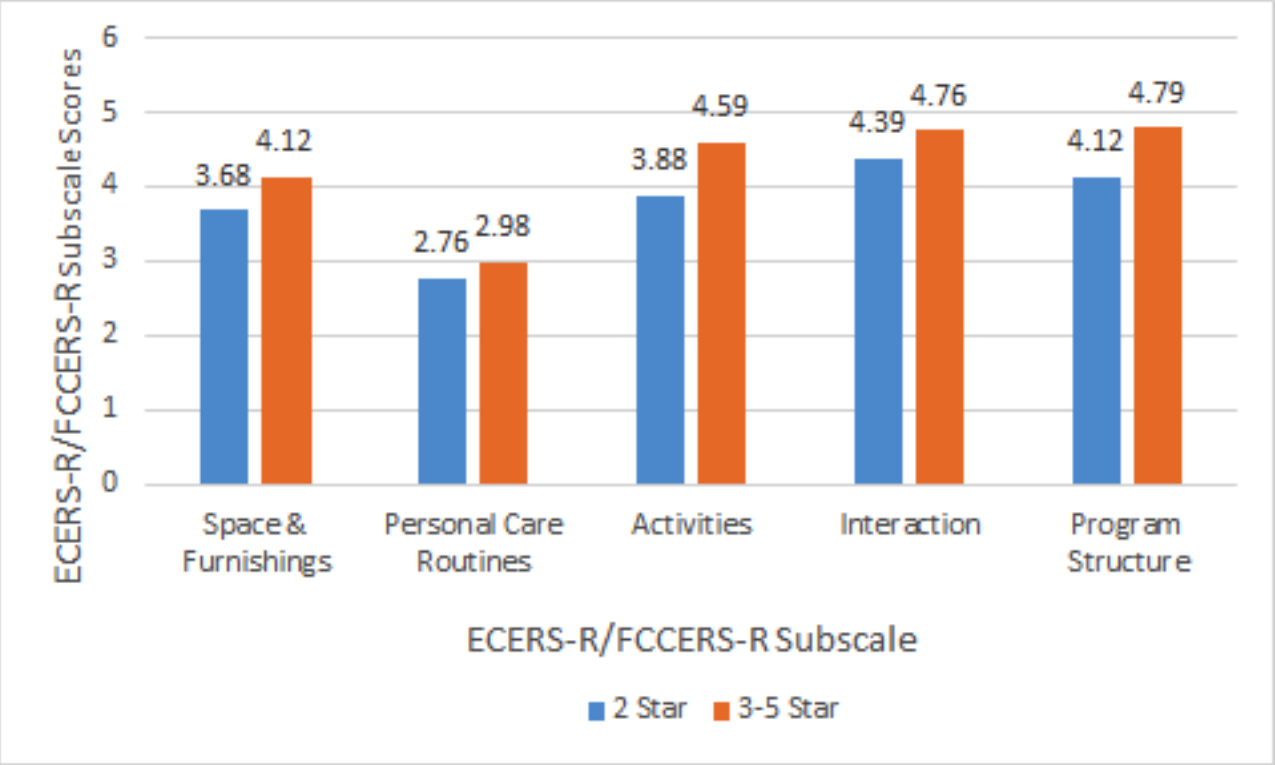


Figure 5: Scores of Language-Reasoning (ECERS-R)/Listening-Talking (FCCERS-R) between Low and High Star Rating

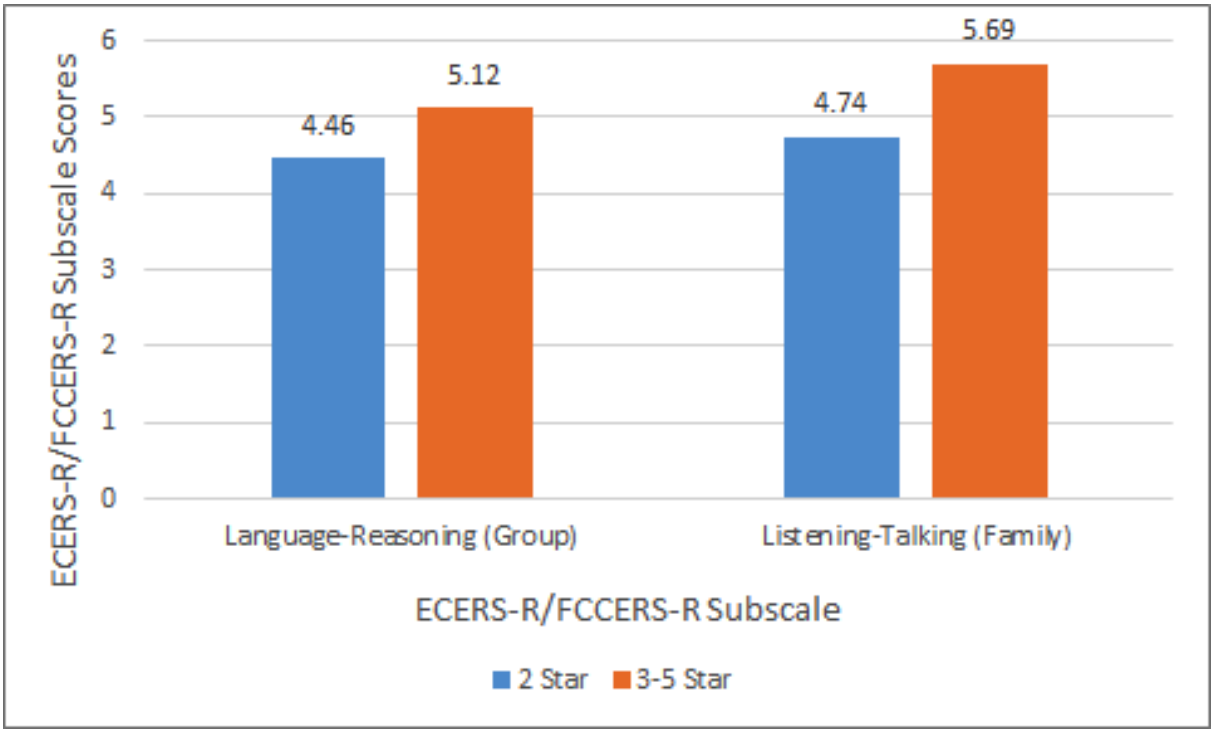


Table 4: Summary of Associations between Observed Quality and YoungStar Rating Level

	Space & Furnishings	Personal Care Routines	Language-Reasoning[a]	Listening-Talking[b]	Activities	Interaction	Program Structure
	b (se)	b (se)	b (se)	b (se)	b (se)	b (se)	b (se)
Model 1: Comparing Low and High Star Ratings							
Low: 2 Star	3.75	2.57	4.69	4.66	4.14	4.59	4.28
	(.12)	(.12)	(.15)	(.40)	(.13)	(.17)	(.18)
High: 3–5 Star	.45**	.20	.65**	.87	.73***	.39	.68**
	(.16)	(.16)	(.20)	(.57)	(.16)	(.23)	(.22)
Model 2: Comparing Star Levels							
2 Star	3.75	2.57	4.69	4.66	4.13	4.59	4.27
	(.12)	(.12)	(.15)	(.41)	(.12)	(.17)	(.18)
3 Star	.24	.20	.71***	.88	.57**	.52*	.63**
	(.16)	(.17)	(.19)	(.59)	(.17)	(.22)	(.23)
4 Star	.92*	.43	.63**	.69	1.00**	.11	.55**
	(.36)	(.34)	(.21)	(.58)	(.33)	(.58)	(.20)
5 Star	1.30***	.13	.41		1.41***	-.12	.95*
	(.25)	(.19)	(.51)	—	(.31)	(.59)	(.38)

Note*:* a. Language-Reasoning sub-scale is only for group providers. b. Listening-Talking sub-scale is only for family providers. All models controlled for region. *p < .05, **p < .01, ***p < .001. The first row under Model 1 and Model 2 shows the regression adjusted average score for the 2 Star classrooms, and the subsequent rows show the regression adjusted difference in score for the relevant higher rated classrooms compared with the 2 Star classrooms. Standard errors of the estimates are in parentheses.

Sensitivity Analyses. As noted earlier, the differential response and recruitment in our sample vary across program type and region, with lower levels of recruitment of family providers and in Milwaukee. We created simple sample weights to correct for differential rates of representation (for cells defined by program type, star rating, and region). When these weights were applied in all analyses, the results were robust, and the differences across star levels were slightly larger. We prefer the unweighted analyses because the weighting scheme was relatively simple, and given that programs are stratified by quality, differential recruitment rates should not, in theory, affect the analyses (especially if we are controlling for region and if ratings do not differ by program type). Empirical results confirm this expectation.

The ERS scores for 10 child care programs were missing a rating on one indicator item related to the quality of supervision of gross motor activities. As noted earlier, this occurred because of confusion about how to code this item if gross motor activities were not observed. In order to see if the omission of this item affected the results, we conducted two additional sets of analyses. First, we excluded the 10 sites with this item missing from the sample, and estimated all the regression models. The pattern of results did not differ with the exclusion of these sites. Next, we also used a best guess for what the rating would have been in instances in which the interviewer provided notes about the teachers' responses to related questions and observations of more general supervision. The validity of these post-hoc ratings is questionable, but it provides some indication of the sensitivity of our results. When regression analyses were conducted with these post-hoc ratings included, the results did not differ from the overall pattern presented in the tables. This provides considerable confidence that one item missing from the scale for ten providers does not affect the results.

Do YoungStar Rating Points Predict ERS Scores?

A second question related to the validity of the YoungStar rating scale is the extent to which the number of points within each rating domain predicts the observed quality of the programs' classrooms. This question is relevant because it is worth understanding whether each domain of rating contributes to the observed associations between star rating and observed quality, or if one rating domain is largely responsible for driving these associations.

YoungStar ratings were provided by Wisconsin's Department of Children and Families for all programs in the study. However, only programs that were given a technical or formal rating had detailed information about how many qualifying points were earned in each rating domain. As a result, the following analyses are based on the 122 programs with a technical or formal rating, in which levels of points were collected. Given that all 5 Star programs in our study achieved their rating by an automated process, this rating level is not represented in these analyses.

As would be expected, summary statistics in Table 5 demonstrate that the total amount of points and all points within specific domains are ordered such that lower star levels have lower points than higher star levels. Because the system has minimum requirements to achieve a higher rating, particularly in the education and training

domain, it would be possible for lower rated programs to have similar or higher points in some domains than more highly rated programs. Yet, the data suggest that, on average, more highly rated programs have more points than low rated programs.

Table 5: YoungStar Rating Points by Star Level (only for Providers with Technical or Formal Rating, N=122)

	September 2013				May 2014			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Total Points								
2 Star	8.6	3.9	2	18	9.5	5.4	0	21
3 Star	18.4	4.1	11	30	18.2	5.4	0	29
4 Star	28.8	4.1	24	35	29.2	5.0	24	37
Education and Training								
2 Star	1.3	2.0	0	9	2.4	2.5	0	10
3 Star	7.3	2.9	3	15	7.1	3.4	0	15
4 Star	12	1.9	10	15	12	1.9	10	15
Learning Environment and Curriculum								
2 Star	2.5	1.0	2	6	2.3	1.3	0	6
3 Star	3.3	1.7	2	8	3.3	1.8	0	8
4 Star	7	2.4	5	11	7.8	3.1	5	13
Business and Professional Practices								
2 Star	3.2	1.8	0	6	3.2	2.1	0	6
3 Star	5.2	1.1	1	7	5.1	1.5	0	7
4 Star	6.2	0.4	6	7	6.2	0.5	6	7
Health and Wellbeing								
2 Star	1.6	1.1	0	5	1.6	1.3	0	5
3 Star	2.5	1.2	1	5	2.7	1.2	0	5
4 Star	3.6	0.9	3	5	3.2	1.3	2	5

Note: Star levels and type of rating process are based on the information from September 2013 for both panels. The total points exceed the maximum points for each star level due to some sites not meeting the minimum requirements or receiving higher points/star levels in May 2014. At both time points there are 49 2 Star, 68 3 Star and 5 4 Star providers represented in the table.

The correlations among the rating scale point domains are provided in Table 6, and these estimates suggest that these domains are highly correlated, which means that programs are likely to have consistently relatively high (or low) point levels across two or more rating domains, although this is not uniformly the case; some programs may have higher (or lower) points in one domain than others. This is important because it

suggests that the rating criterion do not appear to be fully redundant, although they do generally work in the same way. Interestingly, the Child Health and Wellbeing domain is the most highly correlated with all other domains. The lowest correlation is found between Education and Professional Training area and Learning Environment and Curriculum area.

Table 6: Correlations of Rating Points among Four Domains

	September 2013			
	Education	Learning	Business	Health
Education	—			
Learning	.44	—		
Business	.54	.67	—	
Health	.62	.66	.72	—

Note: All correlations are statistically significant at p<.05.

The bivariate regression results summarized in Table 7 assess the extent to which each YoungStar domain of rating points predicts the quality measures, without accounting for any other domain of points. The results indicate that, in general, the total number of points in each domain is strongly predictive of observed quality.[2] It is notable that all other categories of points predict the overall ERS quality score, and at least four of the five subscale scores.

It is also of interest to understand whether the points in particular rating domains are uniquely predictive of observed quality. That is, when other point areas are held constant, do points of a particular rating domain predict observed quality above and beyond other rating domains? The results summarized in Table 8 suggest that two YoungStar point areas seem to be uniquely predictive of observed quality. Business and Professional Practices and to a lesser extent Learning Environment and Curriculum predict the unique variation in ERS quality for both the global measure and its subscales (holding constant rating points in other domains). This does not suggest that the other domain areas, Education and Training and Health and Wellbeing, are not effective as a rating criterion, but it does suggest that the variation that these points predict is also predicted by points in the other rating domains.

[2]However, it is important to note that the magnitude of the associations are not directly comparable because the number of points within the rating areas differ. As a result, domains with a smaller range of points have larger coefficient estimates. The coefficient estimates the expected increase in ERS outcome measures as a result of a one point improvement in the specific rating points.

Table 7: Bivariate Association between for ERS Scores and YoungStar Rating Points (only including sites with Technical or Formal Rating)

September 2013	Overall ERS Scores b (se)	Space & Furnishings b (se)	Personal Care Routines b (se)	Language-Reasoning[a] b (se)	Listening-Talking[b] b (se)	Activities B (se)	Interaction b (se)	Program Structure b (se)
Total Rating Points	.06***	.04***	.03*	.07***	.06	.06***	.07***	.08***
	(.01)	(.01)	(.01)	(.02)	(.03)	(.01)	(.02)	(.02)
Education & Training	.08***	.06*	.06*	.09**	.11	.09***	.09*	.12**
	(.02)	(.02)	(.02)	(.03)	(.05)	(.02)	(.04)	(.03)
Learning & Curriculum	.15***	.15***	.07	.10*	.15	.16***	.18**	.20**
	(.03)	(.04)	(.04)	(.05)	(.13)	(.04)	(.07)	(.06)
Business & Professional Practices	.22***	.16**	.10	.29***	.26*	.25***	.34***	.29**
	(.04)	(.05)	(.06)	(.07)	(.13)	(.05)	(.07)	(.08)
Health & Wellbeing	.20***	.08	.07	.38***	.05	.23**	.33**	.32**
	(.06)	(.07)	(.07)	(.08)	(.21)	(.07)	(.09)	(.10)

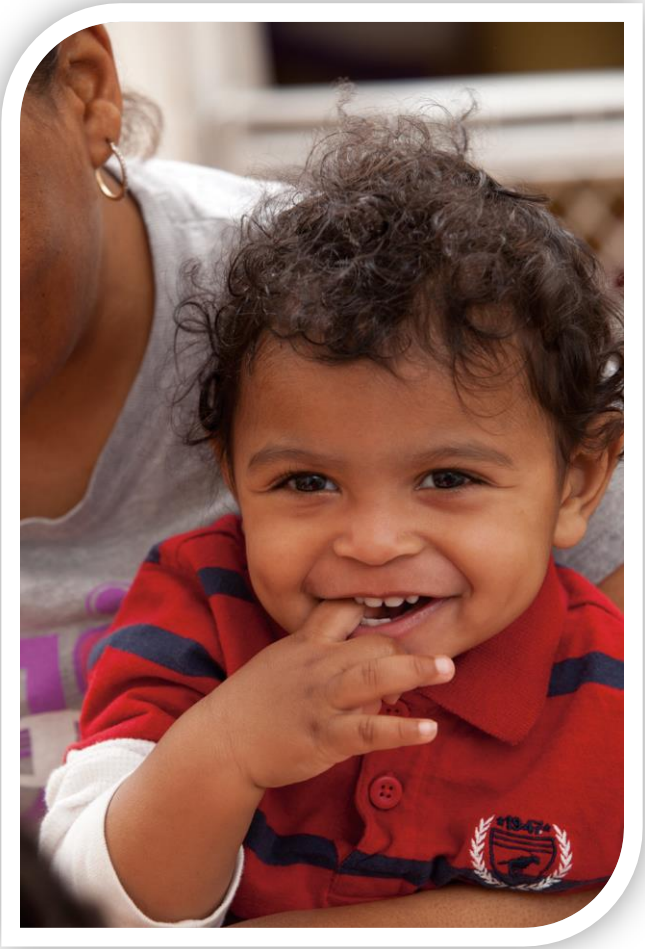
Note: N = 190 classrooms (166 from group sites and 24 from family sites). a. Language-Reasoning sub-scale is only for group providers. b. Listening-Talking sub-scale is only for family providers. Coefficients are from separate bivariate analyses for total rating points and each rating category. *p < .05, **p < .01, ***p < .001. Standard errors of the estimates are in parentheses.

Table 8: Multivariate Association between ERS Scores and YoungStar Rating Points (only including sites with Technical or Formal Rating)

September 2013	Overall ERS Scores b (se)	Space & Furnishings b (se)	Personal Care Routines b (se)	Language-Reasoning[a] b (se)	Listening-Talking[b] b (se)	Activities B (se)	Interaction b (se)	Program Structure b (se)
Education & Training	.04	.04	.06	.02	.09	.04	.01	.07
	(.02)	(.03)	(.03)	(.03)	(.07)	(.03)	(.04)	(.05)
Learning & Curriculum	.08*	.14**	.05	-.03	.07	.09	.06	.09
	(.03)	(.05)	(.05)	(.06)	(.14)	(.04)	(.07)	(.07)
Business & Professional Practices	.16**	.13*	.06	.19*	.21	.18**	.28**	.17
	(.08)	(.06)	(.07)	(.09)	(.16)	(.07)	(.09)	(.10)
Health & Wellbeing	-.04	-.18	-.09	.22	-.31	-.03	.06	.03
	(.07)	(.09)	(.10)	(.14)	(.19)	(.10)	(.11)	(.14)

Note: N = 190 classrooms (166 from group sites and 24 from family sites). a. Language-Reasoning sub-scale is only for group providers. b. Listening-Talking sub-scale is only for family providers. All models include points from all four rating categories. *p < .05, **p < .01, ***p < .001. Standard errors of the estimates are in parentheses.

Sensitivity Analyses. Our analyses used the rating points from September 2013, because these were the points measured at the time the study was fielded. However, over months the points may have changed due to improved program quality or staff turnover. Using the points recorded in May of 2014 (rather than September of 2013) in the analyses yielded findings that were virtually identical to those reported in the tables.



The estimation approach used in the main analyses models all of the rating points as linear measures, which assumes that a one point increase has the same association with improved quality at all points along the distribution. Education and Professional Training has the most possible points, which are largely clustered at the lower end of the distribution, and thus may be most likely for this linear assumption to be violated. As a result, we tested several alternative measures of points in this domain, trying both collapsed categories into an ordinal measure, as well as categorical measures of these collapsed categories, and a log transformation. Overall these other measures did not yield substantively differing patterns, especially when the other rating domain points were held constant.

As was the case for the other analyses, we also estimated models in which we used a weight to account for slight differences in recruitment across region, star rating, and program type. The results were substantively the same when the weights were applied in terms of magnitude and levels of statistical significance. Finally, we also estimated models that tried to test how sensitive our results were to the inclusion of the 10 sites in which the item pertaining to the quality of supervision during gross motor activities was not included in the scores. Again, we estimated models based on our best approximation for what that item would have been scored based on written observation comments for how teachers responded to questions and the scores on other related items, and also excluding these observations all together. The results were robust to both approaches giving us confidence that this missing item was not biasing our results.

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 independently observed measures of classroom quality. The WECCS study was undertaken to provide such an examination of the validity of YoungStar's rating scale in regards to observed quality. Additional reports will examine questions of validity with respect to children's learning and development.

WECCS was successful in recruiting and enrolling a sample of community child care providers in the two selected regions and of varying star ratings and program types. Observational ratings were completed in the winter and early spring of 2014. Results from analyses summarized in this report provide answers to two important questions about the validity of the YoungStar rating scale. First, the data suggest that the star rating level does differentiate among programs of varying observed quality. In particular, programs rated as 2 Star had scores on the global ERS that were about .5 points lower than programs rated at 3 Star or above. These differences were statistically significant and meaningful. Yet, it is important to note that differences reflect improvement within the range of minimal (ERS=3) to good (ERS=5) quality care. Because the ERS ratings were clustered in a fairly narrow range on the scale (standard deviation is about .9), this represents a fairly large proportion of the variation in score ratings. For example, the estimated difference in low vs. higher quality rating translates to effect sizes of about .55 for global quality (suggesting that the difference is over half of a standard deviation, commonly recognized as a substantial effect). The study was not designed to test for differences in observed quality between programs at the higher end of the rating scale.

The secondary validation question answered by this study was whether the underlying rating points were also predictive of a program's observed quality. Data suggested that, as expected, the amount of rating points in each domain was highly correlated and, thus, measure related aspects of program quality. Most importantly, the total number of points in all four rating domains predicted observed quality. When the points for rating domains were considered simultaneously, two domains seemed to demonstrate unique predictive power—Business and Professional Practices and to a lesser extent Learning Environment and Curriculum. With respect to total rating points, the difference in points between a 2 Star program (8.6 average points) and 4 Star program (28.8 average

points) predicts a 1.2 point differences in ERS, which again translates into quite a substantial effect, given the amount of observed variation across programs (over a standard deviation).

There are several key limitations to the current study that should be noted. 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, the study did not have a sufficient number of highly rated programs to consider whether differences in ratings at the high end of the scale were able to effectively differentiate among programs of differing levels of high quality. Third, the study used only the ERS to measure observed quality. The ERS provides an observation of global quality which combines aspects of teacher-child interactions, and space and furnishing with structural aspects of children's experiences in its rating. However, the ERS does not explicitly rate the quality of instructional elements of early care and education nor does it provide a detailed look at the relational quality of caregiver-child interactions. Other observational measures such as the CLASS (La Paro, Pianta, & Stuhlman, 2004) and the Arnett Caregiving Interaction Scale (Arnett, 1989) would provide more detailed evidence about how YoungStar programs differ on these more narrowly defined aspects of programs quality. Likewise, the ERS has been found to have only small predictive power for measures of children's school readiness, and thus it is also important to validate the rating scale with respect to children's improvements in school readiness. Finally, the analyses suggest the validity of the rating scale across both family and group providers. However, the study included fewer family providers by design and therefore a larger group of family providers would be needed to draw definitive conclusions.

Nevertheless, the results of these findings provide the first independent investigation of empirical evidence that observed quality is found to be higher among 3 Star or higher rated programs than for 2 Star programs. It also finds that the quality ratings of most child care providers are in the minimal to good range.

REFERENCES

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology, 10,* 541–522.

Burchinal, M. Magnuson, K., & Powell, D. Children in Early Care and Education (2105). In M. Bornstein & T. Leventhal (eds.), *Handbook of Child Psychology and Developmental Science, Volume 4: Ecological Settings and Processes in Developmental Systems*.

Harms, T., Clifford, R., & Cryer, D. (2005). *Early Childhood Environment Rating Scare-Revised Edition*. NY: New York, Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. (2007). Family *Child Care Environment Rating Scare-Revised Edition*. NY: New York, Teachers College Press.

Hestenes, L. L., Kintner-Duffy, V., Wang, Y. C., La Paro, K., Mims, S. U., Crosby, D., et al. (2015). Comparisons among quality measures in child care settings: Understanding the use of multiple measures in North Carolina's QRIS and their links to social-emotional development in preschool children. *Early Childhood Research Quarterly, 30*, 199–214.

Lahti, M., Elicker, J., Zellman, G., & Fiene, R. (2015). Approaches to validating child care quality rating and improvement systems (QRIS): Results from two states with similar QRIS type designs. *Early Childhood Research Quarterly, 30, Part B*(0), 280–290. doi: http://dx.doi.org/10.1016/j.ecresq.2014.04.005

La Paro, K. M. L., Pianta, R. C., & Stuhlman. M. (2004). The Classroom Assessment Scoring System: Findings from the Prekindergarten Year. *The Elementary School Journal, 104* (5), 409–426.

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

APPENDIX 1 SUMMARY OF DATA COLLECTION

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: 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 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 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 2: Site Recruitment Call Completes and Sampled Sites (Call Completes/Calls sampled)

	2 Star		3–5 Star		
	Family	Group	Family	Group	**Total**
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 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 4: Eligible, Non-Refusal Fielded Sites (fielded sites/site goals)

	2 Star		3–5 Star		
	Family	Group	Family	Group	**Total**
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 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 6: Final Site Completes

	2 Star		3 to 5 Star		
	Family	Group	Family	Group	**Total**
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**