Texas Rising Star

FINDINGS FROM THE

Strengthening Texas Rising Star Implementation Study

Executive Summary

Children's Learning Institute
The University of Texas Health Science Center at Houston
Published October 2019

CHILDREN'S LEARNING INSTITUTE

UTHealth
The University of Texas Health Science Center at Houston

Note: This summary is intended for audiences who prefer a high-level summary of the findings and recommendations from the study. To review the study's analytic approach and to examine specific evidence, please refer to the full report, available at https://twc.texas.gov/partners/texas-rising-star-workgroup.

Overview of Texas Rising Star

The Texas Rising Star (TRS) program is a voluntary, quality-based child care rating and improvement system for child care providers participating in the Texas Workforce Commission's (TWC) subsidized child care program. TRS certification is available to licensed centers and licensed and registered home-based child care providers that meet the certification criteria, as defined by the TRS Certification Guidelines. The TRS program offers three levels of certification (2-star, 3-star, and 4-star) to encourage providers to attain progressively higher levels of quality. Star ratings are tied to enhanced reimbursement rates for children receiving subsidies (minimum of 5% higher, 7% higher, and 9% higher, respectively).

Texas Rising Star Assessment

The TRS Assessment is used by workforce development board and child care contractor staff to assess and provide technical assistance to providers pursuing TRS provider certification and ongoing technical assistance for certified providers. The TRS Certification Guidelines contain criteria for director and staff qualifications and training, caregiver-child interactions, curriculum and lesson planning, planning for special needs and respecting diversity, nutrition, indoor and outdoor environments, and parent involvement and education. Within specific categories, providers are evaluated on:

- required "met" or "not met" items for base certification (i.e., 2-star); and
- points-based items scored on a scale of 0–3 points that may lift a provider to a higher star level (i.e., 3- or 4-star).

Study Aims

In September 2017, TWC partnered with the Children's Learning Institute (CLI) at The University of Texas Health Science Center at Houston (UTHealth) for the Strengthening Texas Rising Star Implementation Study. The goals of the study were:

Aim 1: To examine the reliability of the TRS Assessment. This was the primary aim of data collection and is intended to provide key evidence to support removal or revision of items.

- 1a- To determine within and across category functioning of TRS dichotomous (i.e., met/not met indicators) and points-based items (i.e., 4-point rating scales).
- 1b- To examine inter-rater agreement and reliability within and across TRS categories.
- 1c- To examine the stability of star ratings and caregivers' ratings over time.

Aim 2: To examine indicators of external validity of the TRS Assessment across categories and with other measures of quality and outcomes.

Aim 3: To examine qualitative aspects of implementing TRS Assessment training and data collection to determine the impacts of scoring rules and assessment procedures on reliability and system efficiency.

Study Sample

Our recruitment pool was generated by using Child Care Licensing data and included providers that ranged in urbanicity and socio-economic characteristics from seven counties in the Greater Houston Area (Harris, Galveston, Fort Bend, Brazoria, Waller, Liberty, Chambers) and Dallas county. In order the participate, sites needed at least four classrooms, one per age group: infant, toddler, preschool, and school-aged. This criteria was set to ensure the total study sample would include an acceptable number of classrooms from each age group, and that each facility score could be paired with measures associated with each age group.

Additionally, sites were excluded under any of the following conditions:

- Site was less than one year in operation
- License revoked/suspended in last five years
- Site was included in video samples used to support the development of the TRS Assessment Certification Program.

Recruitment Results

- Total number of sites contacted: 1,227
- Ineligible or no response (e.g., did not return phone calls, line disconnected, etc.): 558
- Total declined: 286
- Total agreed to participate: 169
- Total withdrew: 41
- Final study sample: 128 providers

Sample Classrooms by Socioeconomic Status (SES)

Site recruitment contact lists were structured to ensure providers across all SES levels, closely aligning with the statewide SES range, were contacted about the study. A breakdown of the SES levels for the final sample is included in the following table.

Age Groups	Low	Medium	High	Total
Infant (0-17 months)	58	72	59	189

Age Groups	Low	Medium	High	Total
Toddler (18-35 months)	69	96	82	247
Pre-K (3-5 years)	68	113	99	280
School Age (5-12 years)	44	62	42	148
Total	239	343	282	864

The following executive summary includes a discussion of the results and recommendations produced by the study. We first present category-level results and recommendations, followed by cross-category analysis and recommendations. Please reference the full report and appendix for a complete discussion of the data analysis that supports these conclusions.

Key Statistical Definitions

The following terms are referenced through the analysis and recommendations sections:

Internal consistency: A measure of instrument reliability that determines if items within the same category and subcategories measure the same concepts. Internal consistency values greater than .60 are considered acceptable for research purposes. Values above .90 are considered excellent and are the desired level.

Inter-rater agreement: A measure of rater reliability that indicates the extent to which two people scoring side-by-side are able to reach the same rating.

Generalizability coefficient: A measure of rater reliability that indicates the extent to which a team of raters draw similar conclusions, accounting for differences across the raters and sites assessed.

Normality of score distribution: A method of examining item functioning. Item scores can be normally distributed or skewed (i.e., scores concentrated at the low or high ends). Highly skewed items fail to differentiate quality among providers assessed, which contributes little information to the assessment system and results in missed opportunities to capture rich data.

Results and Category-Level Recommendations

Category 1

DIRECTOR AND STAFF QUALIFICATIONS AND TRAINING

Overview

Category 1 includes items relating to the education, experience, and training of staff, including directors and all caregivers. Category 1 includes a combination of met/not met and points-based items. Many of the items require assessors to collect and combine information about multiple indicators of quality (e.g., several specialized types of training that could satisfy a requirement). This means that while the number of items in this section is brief (see table below), the actual number of indicators an assessor must evaluate is high. For example, category 1 for a licensed center-based provider includes 30 indicators for directors and 41 indicators for caregivers within the items shown in the following table.

Subcategory	Number of Met/Not Met Items	Number of Points-Based Items
Director Qualifications	2	3
Caregiver Qualifications	6	2

Category 1 Study Highlights and Recommendations

- **No center met all category 1 requirements for a 2-star rating. No individual item was scored as met by more than 17% of providers.**

- **Data for a high number of facilities was excluded (i.e., scored "not applicable") across several items.**

  Four items in particular had high rates of exclusion (e.g., 86% excluded for volunteer and substitute caregiver orientation). This suggests these items are not consistently contributing information to provider scores as currently written.

- **Several item-level indicators (i.e., criteria that contribute to item scoring) are difficult to consistently capture based on typical personnel files (i.e., requires information many people do not document).**

These include:

- Years of experience within a TRS or TRS-recognized nationally accredited center
- Years of experience within a licensed or registered child care facility
- Current job status (e.g., difficult to track transitions between full time, part time, substitute, volunteer)

▶ Category 1 is time intensive for assessors to score.

On average, it required 30-40 minutes per caregiver/director for study assessors to review related documents. Record review may approach 90 minutes for early childhood professionals with extensive years of experience and documentation. When the Texas Early Childhood Professional Development System (TECPDS) was used to facilitate scoring, time estimates dropped to 10-15 minutes. The study team developed worksheets that better facilitate scoring of the items, which improved the thoroughness and accuracy of review.

▶ Many of the key elements required for category 1 were more easily scored using TECPDS individual profile reports of staff qualifications and training than direct review of personnel files.

The authors recommend increasing integrity of category 1 scores by relying on TECPDS individual profile reports to reduce scoring errors, ensure authenticity of documents related to staff qualifications and training, and if desired, automate scoring of items based on TECPDS data.

▶ We recommend to revise or remove item-level indicators that:

- have a high rate of N/A scores, unless the indicator is strongly supported by theory and/or evidence;
- do not differentiate provider quality (i.e., highly skewed scores), which will lessen the burden on providers and assessors and reduce the amount of time required to complete an assessment; and
- are inconsistently captured and available for review. Conversely, TRS could set new field expectations and norms for including this information in routine document issuing and management practices.

Category 2

CAREGIVER-CHILD INTERACTIONS

Overview

Category 2 includes items relating to group size, caregiver to child ratio, and the quality of interactions between caregivers and children in the classroom across four subcategories (shown in the following table). Staff ratios and group sizes are structural features of quality but scored as points-based items. The remaining items are process features of quality and are scored as points-based items.

	Number of Items by Age Group			
Subcategory	**Infants**	**Toddlers**	**Preschool**	**School-Age**
Staff Ratios and Group Size	1	1	1	1
Language Facilitation and Support	10	10	10	10
Play-Based Interactions and Guidance	3	3	3	3
Support for Children's Regulation	0	7	7	7
Warm and Responsive Style	6	6	6	6

Category 2 Study Highlights & Recommendations

- **With rigorous training, the assessment team was able to reach reliability for category 2 items.**

- **We examined for differences in scores for the group size/ratio item when using enrollment data (i.e., current scoring criteria) versus staff and children present during the observation period.**

  The latter calculation shows acceptable score distribution and stronger correlations with caregiving behavior. We therefore recommend adjusting the scoring criteria for this item.

- **Several items that rely on frequency counts of behaviors to measure qualitative aspects of caregiving still require a high degree of rater training in order to reliably score.**

  For instance, without training to reliability, assessors are likely to differ in their interpretation of whether or not and how many times a specific behavior is present during an

observation. The study was able to identify alternate scoring that results in reduced ceiling effects and greater reliability for these items. The alternate method scores items based on the caregiver's style (a global rating of the quality and consistency of caregiving behaviors throughout the observation, offset by neutral and harsh negative behaviors) across different settings (e.g., meal time, structured or unstructured activities, and equal engagement with children). We therefore recommend revising the scoring of frequency-based items to align with the alternate scoring criteria.

- **Internal consistency for category 2 for all items using both current and alternate scoring methods is in the excellent range (.90 and above) for all ages.**

Category 3

CURRICULUM

Category 3 includes items broadly related to curriculum, including lesson plans, instructional formats that caregivers use in the classroom, planning for special needs, and considerations for children from bilingual and culturally diverse backgrounds. All items are points-based items.

	Number of Items by Age Group			
Subcategory	**Infants**	**Toddlers**	**Preschool**	**School-age**
Instructional Formats and Approaches to Learning	5	5	5	5
Lesson Plans & Curriculum	4	4	10	1
Planning for Special Needs & Respecting Diversity	3	3	3	3

Category 3 Study Highlights and Recommendations

Category 3 is not functioning well in terms of internal consistency and distribution of scores.

- **All items**

Internal consistency for category 3 for all items using both current and alternate scoring methods is in the borderline acceptable range for infants (.66 and .69, respectively) and toddlers (.60 for both scoring methods). Internal consistency for preschool items reaches the good range for both current and alternate (.85 and .81). School-age internal consistency is unacceptable for both scoring approaches (.51 and .47).

▶ Lesson Planning

- Although preschool items show some signs of reliability, lesson planning items as currently written are not providing a strong measure of curriculum. Substantial conceptual changes to category 3 are recommended to more meaningfully account for curriculum-related practices. Key considerations:
    - The ratings system does not differentiate quality among providers (i.e., highly skewed score distributions).
    - Lesson planning items were among the most difficult to achieve initial reliability for, and the most time-intensive items to score within the assessment, requiring on average 30-45 minutes per classroom for infant, toddler, preschool, and school-age.
    - Given the subjectivity involved in scoring lesson plan alignments based on limited lesson descriptions, the considerable amount of time required to score the items, and lack of evidence to support this approach to measuring curriculum, we recommend removal or substantial revision of lesson plan items. We offer suggestions for more substantive ways to address lesson plans within the TRS system (e.g., score based on observed implementation, process interviews, inclusion in TRS-supported quality improvement plans) in the full report.

▶ Special Needs and Respecting Diversity

These items are too often excluded (i.e., scored N/A) to consistently reflect quality in these areas. We recommend removal or substantial revision of planning for special needs and respecting diversity items as currently measured. We offer suggestions for more substantive ways to address these critical caregiving practices within the TRS system (e.g., process interviews, inclusion in TRS-supported quality improvement plans) in the full report.

▶ Instructional Formats and Approaches to Learning

Given that the items related to instructional formats and approaches to learning (IFAL) are more focused on specific aspects of caregiving behavior, and that scores for these items are more normally distributed, we recommend to move IFAL items to category 2. Correlations between IFAL and category 2 are significant and in the moderate to large range, suggesting they may be appropriately scored together.

Category 4

NUTRITION AND INDOOR / OUTDOOR ENVIRONMENT

Category 4 includes items related to nutrition (policies at the facility level and practices at the classroom level), as well as the equipment, materials, and arrangement of indoor and outdoor learning environments. The nutrition and indoor learning environments subcategories include

a combination of met/not met (required) items and points-based items. The outdoor learning environment subcategory is scored using points-based items only.

	Number of Items by Age Group				# of Met/ Not Met Items (at facility-level)
Subcategory	**Infants**	**Toddlers**	**Preschool**	**School-age**	
Indoor Learning Environment	7	7	7	8	0
Nutrition	3	3	4	3	4
Outdoor Learning Environment	5	4	4	4	0

Category 4 Study Highlights and Recommendations

- **Several items showed limited variation in score, indicating that these items do not differentiate quality among providers.**

  For example, items related to homework practices and meal planning policies and practices showed limited variation. We recommend these items be removed or substantially revised to lessen the burden on providers and assessors and reduce the amount of time required to complete an assessment.

- **The ratings system for nutrition contains too few items to be able to fully assess reliability, and several of these items show limited variation.**

  Removal of low performing nutrition items resulted in improved category 4 reliability. Nutrition practices may be more appropriately captured in a continuous quality improvement framework, as described in recommendation 6.

- **Indoor learning environment items (across all ages) show acceptable reliability.**

- **Outdoor learning environment items show acceptable reliability for all ages except infants.**

- **There were no notable differences in internal consistency for the current and alternate scoring methods.**

Internal consistency for category 4 infant items is borderline acceptable (.60). Toddler, preschool, and school-age items show internal consistency in the acceptable range (.79 to .80).

Category 5

PARENT EDUCATION AND INVOLVEMENT

Category 5 includes items relating to the education and involvement of parents and family members in the program. Both subcategories contain a combination of points-based and met/not met items. Scoring is based on director self-report and document review.

Subcategory	Number of Met / Not Met Items	Number of Points-Based Items
Parent Education	2	2
Parent Involvement	3	3

Category 5 Study Highlights and Recommendations

- **Several of the indicators do not involve objective review of evidence such as documents or observed behavior, and instead rely heavily on self-report.**

- **A few items showed limited variation in score.**

  For example, 96% of providers met S-PE-02, an item related to the school-parent communication system. We recommend removal of S-PE-02 for this reason.

- **Given that the category includes a small number of items, and only acceptable reliability was established, we recommend adjusting the weight of this category within the overall star rating calculation when further validity data becomes available.**

- **Internal consistency is in the borderline acceptable range (.70).**

  Given that items are normally distributed and all items correlate moderately with the total score, the effects of item removal were not examined.

Cross-Category Findings and Recommendations

We made adjustments to categories (e.g., removal of specific items) based on item-level screening procedures (reported in the category highlights) and used factor analysis to confirm the number of underlying constructs within the recommended structure of the assessment. We also compared generalizability coefficients, internal consistency, distribution of star ratings, and stability of ratings over time using the current and recommended structures. Convergence in the evidence across multiple analytical approaches improves our confidence that recommended changes will improve performance of the TRS assessment.

Note: Items in category 1 were not evaluated using measures of internal consistency or factor analysis given that the items were not intended to measure one construct and are based on factual data (e.g., diploma) rather than judgements of quality (e.g., behavioral observation).

Recommended Structure: Confirmatory Factor Analysis

Factor analysis is a statistical method used to explore or confirm the number of underlying constructs (i.e., concepts measured by the TRS assessment) and examine the extent to which the items are designed to measure the same construct. This analysis increases confidence that items within categories measure the constructs the TRS program intends to measure. The confirmatory factor analysis for category 3 indicated a one-factor structure in which lesson planning and curriculum items were measuring one construct in the preschool age group only. Given that instructional formats and approaches to learning (IFAL) items show strong relations with category 2, we included IFAL in the factor structure for category 2. Including IFAL, results for category 2 indicated a one-factor structure fitted data well in four age groups, meaning the final items in category 2 were measuring one general construct. These data support a recommendation to combine these items into a single category representing caregiver-child interactions. For category 4, the results confirmed three separate dimensions exist within this category (i.e., indoor learning environment, outdoor learning environment, nutrition), indicating category 4 does not measure a single construct. For category 5, the results showed a one-factor structure fitted data well, suggesting final items of this category were measuring one construct.

Overall Internal Consistency for Points-Based Items

Although internal consistency was strong using the current structure, we re-examined Cronbach's alpha for points-based items using the recommended structure and found small improvements across all age groups. Additionally, for infant items internal consistency was improved from the "good" to "excellent" range, resulting in "excellent" internal consistency for all age groups.

Inter-Rater Reliability

Generalizability coefficient was estimated for the 10 raters released for independent classroom assessment using the current TRS Assessment structure. G-coefficient was estimated overall for all points-based, classroom-level items in categories 2, 3, and 4 for the current and alternate scoring methods, with rater-level reliability under current scoring ranging from .67 ("marginally acceptable" range) to .89 ("acceptable" range). Generalizability coefficients were slightly higher for the alternate items, ranging from .71 to .89. Of the 10 raters released for independent scoring, reliability for six assessors was calculated as in the "acceptable" range and three in the "relatively acceptable" range. One rater failed to maintain reliability and was reassigned.

We also examined generalizability coefficients under the recommended measure structure (after item removal and confirmatory factor analysis). G-coefficient was estimated overall for all points-based, classroom-level items in categories 2, 3, and 4 for the current and alternate scoring methods, with rater-level reliability ranging from .73 to .90 indicating improved reliability under the recommended structure. This provides evidence to support the use of the new measure structure as a means for improving the accuracy and reliability of field ratings.

Distribution of Star Ratings

In our sample, no providers met all of the requirements for 2-star certification (i.e., met all met/not met indicators). We also examined the percentage of providers with met/not met ratings within categories. Within category 1, no providers met all met/not met items. Within category 4, three providers (2%) met all met/not met items. Within category 5, 23 providers (18%) met all met/not met items. For many items that require providing documentation or self-reporting information that aligns with the TRS standards, it is possible that providers could meet these requirements if standardized templates and sample documents were available.

Because no providers met 2-star requirements, we excluded met/not met items to examine variation in star ratings based on points-based items. The distributions below reflect star rating based on points-based items only, by category.

Category	Number of Providers Per Category Star Rating (**excluding** met/not met indicators)		
	2-Star	3-Star	4-Star
1	115	12	1
2	114	14	0
3	128	0	0
4	110	18	0
5	79	28	21

We also examined the distribution of star ratings under the recommended structure (i.e., excluding items recommended for removal), and found no changes in overall star rating and very few changes within category scores.

Initial Exploration of External Validity

While the primary scope of the study was to examine for and support reliability, where study data allowed, we also examined for relations across categories and among TRS items and external sources that provide initial evidence that TRS scores correlate with other aspects of quality in expected ways. Questions examined include:

Are star ratings stable across brief periods of time?

Stability of ratings was measured by capturing changes in category and overall star ratings in between repeated assessments of the same providers. Ratings stability is important because a single assessment results in a star rating that can be held for up to three years, and star ratings have implications for reimbursements and technical assistance. The study selected 40 facilities and 269 classrooms from the full study sample for participation in the stability rating sub-study. All 40 facilities received two assessments, and a subsample of 16 facilities (n=105 classrooms) received an additional third assessment. On average, assessment 2 occurred 2.5 weeks after assessment 1, and assessment 3 occurred 8.2 weeks after assessment 2.

Change in Star Ratings between Assessments

Overall star ratings were stable across time. It is worth noting that variation in ratings is very limited, with most providers being assessed at the 2-star level. At the category levels, star ratings were also typically stable. Given that the study examined stability over a short length of time, is it recommended to further investigate whether ratings remain stable across the three years of certification.

Stability of Ratings at the Classroom Level

Stability was more of a concern at the classroom level, and in particular within the category 2 (caregiver-child interactions) score used to assign star rating (i.e., the average of median scores across items). Differences in average scores within category 2 from observations 1 to 2 (n=269) and observations 2 to 3 (n=105) were small but statistically significant. Differences in scores for categories 3 and 4 were not statistically significant over time.

Stability of Ratings across Classrooms with Consistent Caregiving Staff

Changes in caregiver were frequent in our sample, even over relatively brief periods of time.

Sixty-six percent of classrooms had a stable lead caregiver across three assessments. Fifty-nine percent of classrooms had stable caregiving staff (including both lead and co-caregivers) between visits 1 and 2. Thirty-eight percent retained the same classroom makeup across three assessments. Although TRS is trying to capture information about children's typical experiences,

it is worth noting that many children in the centers in the study sample are not experiencing continuity of care, which may make it difficult for children to build relationships with individual caregivers.

To learn more about the extent to which the measures themselves show stability when rating the same caregivers across repeated observations, we analyzed stability for a subsample of 40 classrooms where all caregiver assignments were consistent across timepoints. In the subsample of classrooms (n=40) that retained the same classroom makeup (i.e., all caregivers the same across time), there were small but significant decreases in category 2 scores over time. Category 2 primarily measures characteristics of individual caregivers (e.g., warmth and responsiveness). Caregiving behaviors may be higher quality at assessment 1 due to greater motivation on behalf of the caregiver to demonstrate elevated performance at an initial assessment. It is also possible that this effect can be attributed to rater mindset or behavior (i.e., raters may tend to inflate initial rating), despite our intensive efforts to maintain reliability. Consistent with findings from other studies, this suggests that multiple timepoints and raters may be needed to yield a more stable rating of quality at the classroom level. These differences were detected with an average of 2.5 weeks between assessments 1 and 2, and 8.2 weeks between assessments 2 and 3. Further study to learn more about the extent to which these differences relate to other caregiver characteristics and/or children's experiences may be warranted.

Classroom averages for categories 3 and 4 appear to be more stable over time. This may be because some items in these categories are less dependent on individual caregivers and capture the resources and practices of the center (e.g, curriculum, materials, and equipment provided by the director). It is also possible that the items themselves are not as sensitive to changes in practice or the environment as items in category 2.

We re-examined stability across time for all 269 classrooms using the recommended structure and found that the differences for caregiver-child interactions for observations 1 and 2 were still significant, but the differences between observations 2 and 3 (for 105 classrooms) were no longer significant. This suggests that scores are more stable under the recommended structure.

Given that the study examined stability over a short length of time and within a relatively small sample of providers, is it recommended to further investigate whether ratings remain stable across the three years of certification.

Is there evidence that star ratings and classroom quality vary by socioeconomic status?

We explored variation in scores for met/not met indicators based on SES, and found only a few items with identifiable SES differences. It is important to note that most providers, regardless of SES, scored Not Met on most indicators. We also examined for differences in point-based scores. For the current TRS scoring procedure, there is a *slight* trend toward higher scores within higher SES providers. It is important to note, however, that even in the highest rated SES group, providers on average would not meet the threshold for a 3- or 4-star rating at the category level.

Is accreditation related to TRS scores?

TRS providers that are nationally accredited have an opportunity under the current program rules to bypass formal assessment and enter TRS as a 4-star provider. This method for onboarding new providers to QRIS has been used in several states to increase participation under the assumption that standards in place for accreditation are related to QRIS quality standards. Our sample included 18 accredited providers, all of which received a full site assessment. None of these providers scored at the 4-star level on points-based items. Scores for accredited providers were slightly higher than non-accredited providers for categories 2, 4, and 5, but these differences were not substantial enough to change overall star ratings. Based on this sample of providers, we did not find evidence to support automatic 4-star ratings for nationally accredited providers.

Do directors with higher levels of education, training, and experience have higher scores on TRS facility scores?

We examined for correlations between all category 1 director-focused items and TRS classroom items and found no consistent patterns. Given that TRS qualifications items are scored based on combinations of many indicators, we also looked at the extent to which individual indicators (e.g., years of experience, business management training hours) relate to classroom and facility points at the category level. We found multiple small to moderate significant correlations with facility-focused categories. This suggests information is lost with the current item structure, which may limit predictive validity.

Do caregivers with higher levels of education, training, and experience have higher scores in caregiving behaviors?

We examined correlations between all category 1 caregiver-focused items and TRS classroom items and found a fairly consistent pattern of correlations that suggests:

- Providers with more qualified staff (measured by P-CQT-01) have small to moderate correlations with higher scores for category 2 and category 4, and higher category 4 star ratings.
- Caregiver staff training topic alignment (measured by P-CQT-03) is moderately related to category 3 scores.

Do lower caregiver-child ratios relate to higher TRS scores?

Low caregiver-child ratios are widely considered to be an important structural feature of quality programs that allows caregivers to better supervise children and engage in more positive interactions. In the study sample, better scores for TRS group/ratio shows significant small correlations with category 2 and 4 scores. We also looked to see if ratio relations were stronger for certain age groups within each category and found that correlations were small across all age groups.

Do TRS scores for caregiving behavior (e.g., category 2) relate to another established measure of caregiving quality (convergent validity)?

We examined for evidence of convergent validity by comparing TRS scores for caregiver-child interactions with scores from another established measure of caregiver interaction quality, the Arnett Caregiver Interaction Scale. Multiple highly significant correlations were found with category 2 scores and Instructional Formats and Approaches to Learning (in category 3) than with non-behavioral items. These data provide initial evidence that the behavioral observation items within the TRS assessment relate well to other measures in routine use.

Is the TRS assessment sensitive to changes in caregiver-child interaction quality associated with quality improvement efforts?

We also examined for evidence of TRS category 2 (caregiver-child interactions) external validity in the context of a random assignment pilot study. The pilot study was the initial evaluation of an educational intervention developed to support childcare providers, CIRCLE Infant & Toddler Teacher Training: Play with Me. The pilot sample included 38 teachers in Dallas and Houston (18=intervention, 20=control). Participating pilot teachers had an average of six children per classroom, ages 24-36 months. The total number of children participating in the study was 241 (115 control and 126 intervention). Intervention teachers received the intervention for approximately 6 months.

Controlling for demographic characteristics, we found initial evidence of external validity when examining for growth in category 2 scores. Caregivers in the intervention group showed greater gains in caregiver-child interactions than teachers in the control group, gaining on average 12 points more in caregiver-child interactions scores. This represents a significant improvement in scores and provides evidence that the caregiver-child interactions scale is sensitive enough to detect changes in quality associated with brief quality improvement efforts that are well-aligned with the TRS standards. Changes at the sub-category level ranged from an effect size of .41 to .65. However, changes were only statistically significant for language facilitation and support and warm and responsive style. This finding is consistent with the primary focus of the intervention (i.e., session content was closely focused on language facilitation and support and responsive caregiving).

Is there evidence that children's outcomes are shaped by qualities measured by TRS (external validity)?

Using the same sample described above, we found significant moderation effects of baseline classroom quality for BITSEA Social Competence Total score (b=-.046, p=.025). The results indicated that the intervention works better for students who were in low-quality classrooms at baseline. Moreover, children in the treatment group showed significantly greater gains in social competence than those in the control group when caregiver-child interactions scores were average or below average (ie., 1 standard deviation below mean). This provides additional

evidence that the caregiver-child interactions construct is adequately sensitive to differentiate intervention effects related to caregiving quality.

Study Limitations

This study took place in licensed center-based child care facilities that served all ages. Therefore, the findings presented are not necessarily representative of centers that serve a limited age population (e.g., school-age only) or home-based child care providers. However, given the large sample sizes obtained within each age group, the classroom-level analyses likely generalize to center-based facilities that serve fewer age groups. Given the differences for home-based providers in staffing patterns, child age-group and classroom makeup, and TRS items related to this setting, it is recommended to separately study reliability and validity in home-based child care.

Finally, our initial exploration of validity was limited given our primary focus on reliability and the development of certification procedures to ensure accurate and consistent statewide ratings. Once field reliability is well-established, we recommend the collection of more extensive and diverse validity evidence (e.g., child and provider outcomes) than what was possible in the scope of this study.

Key Recommendations

We recognize that there are multiple goals for Quality Rating and Improvement Systems, including advancing:

- **A market-based system for improving quality** that makes quality transparent to families so that they can make informed choices about where to enroll their children. From this perspective, a QRIS system prioritizes aspects of quality most closely connected with child outcomes and family satisfaction.
- **Workforce professionalization** to improve the level of education and experience of the early childhood workforce, build a stronger sense of attachment and recognition within the profession, improve compensation, and recruit and retain highly qualified staff.
- **Support for child care providers** that demonstrate a commitment to delivering high quality care and improving their services by providing increased financial incentives and targeted technical assistance.

The recommendations provided below may at times differentially serve these goals, and should be viewed through these sometimes competing lenses. For example, items related to Director Qualifications may not be highly correlated with children's classroom experiences, but may be important for promoting the professionalization of the workforce.

★ Recommendation 1: Removing or adjusting low-performing items to improve instrument functioning.

We are recommending retention of the current standards for approximately 71% of the items in the TRS assessment. Of the items recommended for retention, we recommend specific revisions to the scoring criteria and/or updates to the technical scoring manual (TSM) for approximately 35 items. We successfully tested alternate scoring for many of these items. We also recommend specific minor TSM updates only for an additional 10 items.

For the remaining 29% of items, we recommend removal or substantial revision of the standard itself or the current scoring approach based on data analysis results, implementation concerns, or both. Data-based concerns for item removal/revision are discussed in the previous sections. Implementation concerns include lengthy scoring times, inconsistent access to required data elements, highly subjective scoring criteria (reliance on provider self-report), and overlap with licensing data. Item removal/revision recommendations are primarily related to lesson planning, nutrition, indoor learning environments, and parent education. We tested alternate items within category 3 but the alternates did not strengthen instrument functioning sufficiently to recommend their use.

Evaluating and measuring curricula continues to be a challenge within the early childhood landscape. Aspects of curriculum implementation may be more fully captured (and provide a more meaningful score of curriculum quality) through a combination of school leader and staff interview protocols, document review, and observations of lesson implementation and student learning. These assessment approaches can be used to establish qualitative scores that can guide quality improvement planning.

The scope of the current study did not include developing and testing new items (i.e., standards) outside the current TRS guidelines. Therefore, evaluation approaches and item recommendations are focused on revisions to the current, TWC-adopted program guidelines to strengthen the TRS assessment.

★ Recommendation 2: Adjusting the relative weight of categories to be more in line with measure reliability and to more accurately reflect the influence of evidence-based practice on children's outcomes.

The current TRS system has five categories that receive equal weight in star rating calculation, regardless of the number of items (e.g., category 2 includes 27 items and category 5 includes 5 items). Thus, the current scoring approach signals equal importance for all categories of quality. While measurement of child outcomes was beyond the scope of the current study, the evidence base suggests constructs aligned with some TRS categories are more closely related to children's experiences and outcomes. For example, there is substantial research evidence that demonstrates that high quality learning experiences within warm and responsive relationships with adults is the best way to advance social-emotional, language, early literacy, and math outcomes for children. These *process* features of care are consistently found to be better predictors of student outcomes than *structural* features of care, such as director qualifications. Because one of the goals of Texas Rising Star is to provide families with clear and accurate indicators of quality, we recommend these aspects of care feature prominently in the quality rating. Within the TRS assessment, caregiver-child interactions (category 2) and instructional formats and approaches to

learning (subcategory of category 3) are highly aligned with the process features prior research has identified lead to better child outcomes. Given that these items have also performed well during the study (e.g., have good internal consistency, relate to other measures of quality), it is recommended that these items should be the most heavily weighted when producing a star-level rating.

There are multiple approaches for adjusting the relative weight. For example, TRS could assign differential weights to each category to align with the evidence base (e.g., category 2 would receive more weight than category 5). An alternative would be to calculate average scores across all items in the recommended assessment structure, which would place more weight on caregiver-child interactions because of the higher number of items measuring this construct. In the long term, our recommendation is to first establish statewide reliability using the recommended structure, followed by a validity study that captures key outcomes aligned with TRS goals (e.g., gains in child skills and financial stability for providers). The results of predictive analysis would be used to guide category weighting decisions, such that categories with low predictive validity across outcomes would receive less weight.

★ Recommendation 3: Revising procedures for automatic certification of nationally accredited providers.

Of the accredited programs assessed, none were scored at a 4-star level. This data suggests that TRS should discontinue automatic 4-star ratings for nationally accredited providers and conduct full site assessments prior to certification. This will also aid targeting efforts in continuous improvement plans.

★ Recommendation 4: Implementing a rigorous training and reliability monitoring process to ensure accurate star ratings across the state.

Inter-rater reliability has significant implications for the fairness of quality ratings attributed to providers and the accuracy of ratings communicated to families. The assessors for this study were able to reach "acceptable" inter-rater reliability after a rigorous training process (see full report for detailed description of training procedures). To ensure accurate ratings across the state, the authors recommend that TRS adopt a similarly rigorous training process using research-supported standards and procedures to reach reliability prior to official data collection. To further strengthen reliability, TRS should consider requiring assessors to be accountable to a central body that certifies reliability and conducts routine reliability monitoring. Given the dispersion of assessment staff across a large and diverse state, centralizing reliability certification and monitoring is the best way to ensure assessment approaches remain aligned, and consequently, that ratings remain fair and accurate representations of quality.

With any instrument, maintaining reliability requires frequent and consistent use; therefore, we recommend that TRS assessors be required to maintain a monthly minimum of classroom observations (e.g., 25 classroom observations per month). Study assessors on average completed 36 classroom assessments per month, with a recommended maximum of three per day. We also recommend establishing monitoring procedures to capture assessor "drift" and prompt re-training efforts when required. Finally, requiring specific notetaking and documentation procedures can help bring clarity to the ratings process, strengthen the accuracy of ratings, and provide evidence for specific scores in communications with providers.

★ Recommendation 5: Standardizing application and scoring routines to improve program efficiency and accuracy of star assignment.

Based on our experiences with data collection for the study, we have identified multiple strategies for streamlining the efficiency and accuracy of ratings, particularly for items that require document review. We recommend to:

- Require specific notetaking and documentation procedures to strengthen ratings and communication about scores with TRS staff and providers.
- Require assessors to utilize TECPDS reports to facilitate scoring of director and caregiver qualifications. See the full study report for a list of indicators that can be captured using TECPDS.
- Integrate TECPDS with the TRS Online Assessment Tool, enabling automated scoring of director and caregiver qualifications.

★ Recommendation 6: Establishing a continuous quality improvement (CQI) framework that uses a developmental approach to ensure providers receive technical assistance and professional development in alignment with their current star ratings.

A CQI approach can be used to target early technical assistance (i.e., before certification) to providers who are not able to meet TRS standards in order to *lift quality and increase participation* in the program. Moreover, the results of this study strongly suggest technical assistance is required to move existing certified providers to *progressively higher* levels of quality that fully meet TRS expectations. Providers delivering high quality services also may need technical assistance in specific areas to *maintain* quality (e.g., after staff turnover). Therefore, we recommend leveraging TRS mentoring staff to provide intensive and individualized technical assistance to achieve these aims. A coordinated CQI framework can include a combination of self-study materials, professional learning communities (PLCs), and individualized coaching. Please view the full report for how CQI activities can be tailored to providers level of quality and stage of TRS implementation.

★ Recommendation 7: Continuing exploration of external validity.

This study focused on strengthening the reliability of TRS ratings to ensure reimbursement rates are accurately allocated and technical assistance is appropriately targeted to needs. The study found some initial evidence of validity (e.g., strong correlations between TRS caregiver-child interactions and validated measures of caregiving quality). Once field reliability is established using the recommended structure, additional research is recommended to further examine long-term rating stability, the ability of the CQI approach to increase TRS participation and advance providers to increasing levels of quality, and evidence that TRS program participation predicts outcomes of interest (e.g, a market-based system for improving quality, workforce professionalization, and support for child care providers).

For more information on the study results and recommendations, please contact ms.cli@uth.tmc.edu.