Texas Rising Star

FINDINGS FROM THE

Strengthening Texas Rising Star Implementation Study

Final Report

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

Table of Contents

Section 1

Introduction

What is Texas Rising Star?

The Texas Rising Star (TRS) program is a voluntary, quality-based child care rating and improvement system of child care providers participating in the Texas Workforce Commission's (TWC) subsidized child care program. TRS provider certification is available to licensed centers and licensed and registered home-based child care providers that meet the certification criteria. 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).

In recent years, many states have adopted quality rating and improvement systems (QRIS) to measure the quality of child care programs and to provide professional development to help these programs improve the quality of care they offer to children and families.

The National Center on Early Childhood Quality Assurance (2013) defines QRIS as "a systemic approach to assess, improve, and communicate the level of quality in early and school-age care and education programs. Similar to rating systems for restaurants and hotels, QRIS award quality ratings to early and school-age care and education programs that meet a set of defined program standards. By participating in their State's QRIS, early and school-age care providers embark on a path of continuous quality improvement. Even providers that have met the standards of the lowest QRIS levels have achieved a level of quality that is beyond the minimum requirements to operate (p.1)."

Across Texas, many parents and families choose to enroll their children into child care programs, including center-based and home-based programs. Numerous research studies have shown that at-risk children who attend higher quality child care programs are more prepared for school entry than children who do not attend quality child care programs (Adams, Zaslow, & Tout, 2007; Booth & Kelly, 2002; Burchinal & Cryer, 2003; Fontaine, Torre, & Grawfwallner, 2006).

Those providers that voluntarily achieve TRS provider certification, offering quality care that exceeds the Texas Health and Human Services Commission (HHSC) minimum Child Care Licensing (CCL) standards for director and staff qualifications, caregiver-child interactions, age-appropriate curricula and activities, nutrition and indoor/outdoor activities, and parent involvement and education, are in a better position to contribute to the early development of

children. As providers progress through the levels of TRS provider certification, they contribute progressively more to the development of the children they serve daily.

History of Texas Rising Star

In the mid- to late-1970s, federal standards for quality child care were implemented across the nation. By the early 1980s these federal standards were discontinued. However, in Texas a state workgroup was then formed to develop standards for child care providers. The research from this workgroup formed the basis for the refinement and development of the TRS provider certification criteria. These criteria were in use from June 1991 to October 2000.

The TRS Child Care Provider Certification Guidelines (TRS Provider Guidelines) were revised and issued in October 2000, incorporating the recommendations of a workgroup formed in 1999. The workgroup consisted of TWC staff, LWDB staff, child care contractors, and child care providers from across the state. In 2000, the revisions mainly updated the assessment and certification procedures. In 2003, TWC updated the recertification and monitoring time frames for TRS providers.

In January 2013, the Texas Early Learning Council (TELC) released recommendations for the state to develop a statewide, cross-sector QRIS for Texas. One of the recommendations included Texas Rising Star as the basis for a QRIS in Texas, influencing the TRS workgroup convened later that year to recommend revisions to TRS.

Effective September 1, 2013, House Bill (HB) 376, 83rd Texas Legislature (Regular Session), amended Chapter 2308 of the Texas Government Code relating to the TRS program. As amended, Chapter 2308 required TWC's three-member Commission (Commission) to:

- Create a TRS program review workgroup to recommend revisions to the TRS program
- Propose rules that incorporate the TRS workgroup's recommended revisions
- Establish graduated reimbursement rates for TRS providers
- Require Local Workforce Development Boards (LWDB) to use at least 2 percent of their annual allocations for quality child care initiatives
- Make funds available for LWDBs to hire TRS Assessors and mentors to provide TRS program technical assistance to child care providers

In 2013, TWC convened a workgroup dedicated to the revision of TRS as required by House Bill (HB) 376 of the 83rd Texas Legislature. The purpose of the TRS workgroup was to recommend revisions to the TRS program. The TRS workgroup invited stakeholders from around Texas to participate in workgroup discussions and provide input into the proposed TRS program revisions. Stakeholders included staff from state agencies responsible for child care program implementation and regulation, LWDB representatives, child care providers, early childhood development experts, advocates and policy makers, and families.

HB 376 required that the workgroup submit recommendations proposing changes to TRS by May 2014, and rules that incorporate the proposed changes by September 2014. The proposed changes to TRS were approved by TWC on January 27, 2015. All TRS providers were certified under the revised guidelines by September 1, 2015.

During the fall of 2015, TWC held several public meetings to solicit input on the child care program, including the TRS program. In January 2016, TWC hosted two provider workgroup sessions and a TRS Assessor/mentor group to gather feedback and recommendations on the 2015 TRS revisions. Based on the input from these stakeholder meetings, the Commission recommended modifications to the TRS Provider Guidelines designed to streamline the application and assessment process and to clarify and improve the TRS criteria. In 2018, the TWC hosted sessions across the state to elicit feedback from providers and TRS staff. Based on the feedback, recommendations were made and went into effect January 1, 2019. TWC continually monitors the progress of TRS and reviews the program every four years.

Strengthening Texas Rising Star Implementation Study

In September 2017, TWC partnered with the Children's Learning Institute (CLI) at The University of Texas Health Science Center at Houston (UTHealth), the designated State Center for Early Childhood Development, to strengthen implementation of the QRIS through an implementation study that focused on three broad initiatives:

- Study the reliability and validity of the TRS assessment system and make recommendations for improvement;
- Develop a sustainable certification and training system for TRS Assessors and mentors to ensure ratings are consistent across LWDB areas and assessors; and
- Test delivery of mentoring protocols aligned with TRS standards, enhancing TRS's Quality Improvement (QI) capabilities.

Current Training Protocols for TRS Assessors

For two years after the initial roll out of the new assessment system, staff attended a 5-day, face-to-face training. These trainings were largely focused on helping staff understand the new TRS standards and assessment procedures, and although trainees spent time practicing scoring with feedback, they were not trained to meet a specific reliability standard. Assessors also had access to ongoing technical assistance from the TWC and CLI to support them with rating accuracy and adherence to assessment procedures (e.g., direct email to TWC TRS specialists, help ticketing on CLI Engage, moderated TRS assessment discussion board, online training course that contains content and exemplars from face-to-face training, annual TRS regional trainings). Questions received through **direct email** via the TWC workgroup email account are those related to protocol, procedures, or implementation of the TRS assessment tool. The **help ticketing system** on CLI Engage functions as the platform to which assessors and mentors can submit and receive guidance related to utilizing the online assessment tool. The **TRS discussion board** is also housed on CLI Engage and is only available to be viewed by LWDB staff. The discussion board functions as a platform for LWDB staff to view questions that were previously asked in various training settings or submitted through email that relate to the protocol and procedures of conducting and completing assessments. The guidance is provided by TWC staff and serves as an additional point of reference for LWDB staff. The TRS online course provides content specific information related to caregiver behaviors and provides an opportunity for LWDB staff to see examples of high quality interactions. The online course also provides pertinent information

related to best practices regarding assessments and gives assessors the opportunity to practice on focusing their observations on behaviors related to specific measures. In 2019, CLI had the opportunity to join TWC staff in five face-to-face **regional trainings** that provided LWDB staff the opportunity to receive training on the specific resources including the TRS assessment tool, resources to support providers in quality improvement efforts, TECPDS resources, and guidance and resources related to mentoring.

Within the current training model, local LWDBs and their contractors are responsible for ensuring new staff are well trained (i.e., onboarding of new staff that occurs in between statewide trainings) and that existing staff continue to perform in alignment with TRS standards and procedures (i.e., long-term adherence to trained protocol). This decentralized approach increases the risk that TRS ratings will vary by region, as team members in close contact with each other start to coalesce around local interpretations and practices.

The Strengthening Texas Rising Star Implementation study and assessment certification system's design provides a scalable approach that ensures all staff are trained to reliability prior to data collection, and includes systems for monitoring reliability and preventing drift among field raters over time.

Measurement Study Aims

▶ Aim 1: To examine the reliability of the TRS assessment.

This is the primary aim of our data collection, and is intended to provide key evidence to support removal or revision of measures. The results of these analyses also informed the development of the training and certification program (discussed in Appendix 9).

1a. To determine within and across category functioning of TRS dichotomous (i.e., met/not met indicators) and points-based measures (i.e., 4-point rating scales). Key questions include:

- Which items are not contributing information that helps to differentiate quality among providers, often referred to as floor (i.e., almost all scores are low, 0 points) and ceiling effects (i.e., almost all scores are high, 3 points)?
- Are there items that are so frequently excluded from scoring (i.e., item not applicable, score N/A) as to cause concern about how and when providers' scores are impacted by the measure?
- To what extent do measures within a particular category or subcategory relate to each other, providing evidence that TRS is measuring what it intends to measure within each of the assessment's conceptual areas (e.g., caregiver child interactions items measure something distinct from those in curriculum)?
- Can the reliability of ratings, within and across categories, be improved through the removal of items or by using alternate scoring criteria?

1b. To examine inter-rater agreement and reliability within and across TRS categories. Key questions include:

- To what extent do raters (i.e., assessors) have reasonable agreement in rating the same provider and caregivers?
- Is inter-rater agreement acceptable given variation in quality among participating providers?

1c. To examine the stability of star ratings and caregivers' ratings over time. Key questions include:

- Is a provider's star rating stable across brief periods of time (e.g., if a program is rated as a 3 star in January are they reassessed as 3 star 1 month later)?
- Are individual classroom and caregiver's ratings stable across brief periods of time (e.g., if a caregiver's interactions with children are rated as high during an observation at the beginning of the month, are they also rated as high 3 weeks later)?

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

Key questions include:

- How does the distribution of TRS scores by category and overall scores vary by regional differences in socioeconomic status?
- Are some measures within the TRS assessment more challenging than others for providers?
- To what extent do structural characteristics of providers and staff relate to process features of care (e.g., do lower caregiver-child ratios relate to higher quality caregiver-child interactions)?
- Do measures across categories relate to each other in expected ways (e.g., do caregivers with high scores on curriculum have higher scores on language support)?
- Is national accreditation status related to TRS provider certification scores (e.g., if a center is NAEYC accredited do they meet TRS standards and at what star level)?
- Do TRS measures relate to teacher and child outcomes in expected ways (e.g., do higher ratings on warm and responsive behaviors relate to children's gains in social skills)?

▶ 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?

Key questions include:

- Can note-taking and documentation be standardized to support rater agreement and discussion of key evidence?
- Are there implementation barriers to accurate data collection and scoring within and across TRS categories? Are there best practices or adjustments to measures or scoring protocols that would improve efficiency in data collection or accuracy of scoring?

Measures and Data Sources Used in the Study

Texas Rising Star Assessment

The Texas Rising Star Provider Certification Guidelines are used by Workforce Development Board and child care contractor staff to assess and provide technical assistance to providers pursuing Texas Rising Star provider certification. The certification guidelines contain criteria for director and staff qualifications and training, caregiver-child interactions, curriculum, nutrition and indoor/outdoor activities, and parent involvement and education.

Each category of the certification criteria is given a star level rating based on the average score across the median values for all points-based measures in that category. A provider's overall star designation is based on the lowest star level achieved across the five categories. The rationale for this scoring protocol is to ensure the provider meets higher quality standards across measures in all categories. An exception is made for providers that receive a four star rating in four of five categories and a three star in the remaining category. These providers receive a four star rating. Providers are evaluated on items across the five categories, with items scored at the facility and class levels by age group. Within specific categories, providers are evaluated on:

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

TRS Assessors utilize several standardized forms to collect information at the facility and classroom levels. The Classroom Assessment Record Form (CARF) includes class-level information collected by TRS Assessors during assessment visits. Individual forms are available for recording information related to classrooms serving specific age groups (infant, toddler, preschool, all ages). The Facility Assessment Record Form (FARF) includes facility-level data collection. Individual forms are available for the various program settings eligible to participate in TRS (center-based, home-based, and school-age programs).

For clarity throughout the report we define assessment information as follows:

- **Assessment** refers to all measures and items captured by TRS (i.e., all 5 categories together)
- **Measure** refers to a category of items (e.g., category 1 Director and Staff Qualifications and Training)
- **Item** refers to an individually scored/rated statement within categories (e.g., item S-DQT-02, Director Training)
- **Indicator** refers to any specific evidence within an item (e.g., specific criteria considered for scoring one item. For example, item P-DEQT-01 includes indicators for college credit hours, credentials/certificates, degrees, and years of experience as a director.)

The following tables show the total number of assessment items by form type and age group. Additional information related to the specific items for each category is provided in the Results section of the report.

TRS Classroom-Level Assessment Total Number of Items by Age Group

Infants	Toddlers	Preschool	School-age
47	53	60	51

TRS Facility-Level Items

Number of Points-Based Items	Number of Met/Not Met Items
10	17

Additional data sources and measures used to support study objectives, including the Texas Early Childhood Professional Development System (TECPDS), Child Care Licensing Daycare and Residential Operations Data, the Arnett Caregiver Interaction Scale, and the Brief Infant Toddler Social Emotional Assessment (BITSEA). These are described below.

The Texas Early Childhood Professional Development System (TECPDS)

Part of TECPDS, the Texas Workforce Registry database contains professional development records voluntarily uploaded by users in all early childhood sectors. Users can upload their professional development, education, and work history into the Texas Workforce Registry in three areas on the website after logging into their accounts. Any updates or changes in a user's information and records are reflected across sections of their account (ensuring all information uploaded into TECPDS is stored in the Texas Workforce Registry database). In order to facilitate scoring of category 1 items related to staff education, training, and experience, study staff worked with providers to load assessment relevant records into TECPDS, which allowed for record validation (e.g., verification of training certificate authenticity) and faster scoring of items within the category. More information about the recommended use of TECPDS in the TRS assessment can be found in the Recommendations section (Recommendation 5).

Child Care Licensing Daycare and Residential Operations Data

Child Care Licensing data for all child care operations in Texas, including TRS providers, was collected from the Texas Open Data Portal (available at data.texas.gov). The portal provides free public access to openly available data across the state agencies. This dataset contains detailed information about all child care programs regulated by the state's Minimum Licensing Standards, managed by the Child Care Licensing division of the Texas Health and Human Services. The dataset includes 38 unique fields that detail information about each provider, including name, address, operation type, programs/services offered, contact information, hours of operation, total capacity, age groups served, and information on deficiencies and reports.

This data was used to create an automated TRS-aligned report detailing the current licensing deficiencies and status changes for TRS providers participating in the study. We used this report in the study to learn more about the extent to which providers were able to meet TRS-selected

licensing criteria required for initial program eligibility and retention of star level/participation. Details of this analysis are shown in the Final Sample Characteristics section.

Arnett Caregiver Interaction Scale

The Arnett Caregiver Interaction Scale contains 26 items designed to measure the emotional tone, discipline style, and responsiveness of a caregiver. The items are organized into the following subcategories: sensitivity, harshness, detachment, and permissiveness. (Massachusetts Department of Early Education and Care, 2011). The Arnett CIS drew from a well-established theory of parenting, linking caregiver interactions to child outcomes in cognitive and socio-emotional development (Colwell, Gordon, Fujimoto, Kaestner, & Korenman, 2013). The scoring ranges from 1-*not at all* to 4-*very much* for the items within this measure. This scale can be used to score multiple caregivers in a classroom separately, showing the variability in styles within one classroom. In this study, the measure is used to explore external validity.

Brief Infant Toddler Social Emotional Assessment (BITSEA)

The BITSEA is a 42-item screener for social emotional/behavioral problems and delays in competence for children ages 12 to 36 months (Briggs-Gowan et al., 2004). The intraclass correlation coefficient for the BITSEA is .87. In this study, the measure is used to explore external validity.

Section 2

Methods

Training Procedures

Recruitment of Staff

Our assessment team was comprised of several existing CLI staff members who had developed and delivered TRS assessment training to TRS staff during two prior required state sponsored trainings. Three of these team members served as master raters as they had consistently demonstrated good agreement with each other and with lead staff from the Texas Workforce Commission's TRS program team. Additional trainees consisted of existing and newly hired research assistants and education outreach mentors, all of whom had either prior early childhood assessment experience with CLI or who had previous TRS program experience with a local workforce development board's TRS contractor.

TRS Assessment Training

Throughout the project, we trained 14 observers using two training models. Given that our goal by the end of the study was to deploy an online training and certification system for TRS staff, we used our training opportunities within the reliability study to iteratively test components of what would become the online system. Initially, our goal was to train all raters to reliability on both the Classroom Assessment Rating Form (i.e., items scored in each classroom) and the Facility Assessment Rating Form (i.e., items score once at the facility level). Of the 16 assessors that were trained, only seven became reliable on CARF assessments and five became reliable on FARF assessments.

In this section we describe the progression of iterative design stages used to train study staff to reliability and to test our approach prior to building the online state training system.

Phase 1 training included 3 major components: 1) understanding TRS program standards and guidelines, 2) building foundational early childhood education and care content knowledge, and 3) TRS scoring practice with feedback and certification.

Component 1: Training closely resembled the prior statewide training approach, consisting of 5 days of face-to-face training that focused on building shared understanding of TRS standards (e.g., scenario analysis, viewing and discussion of video exemplars), assessment procedures and observational protocol (e.g., required observation length and termination rules), and practice calibrating ratings (i.e., mock analysis of authentic artifacts and video recordings of classroom interactions) with feedback from master raters. We used the initial training phase to identify areas of concern for rater agreement and to document examples and non-examples of key behaviors and evidence for inclusion as clarifying information in a study version of the technical scoring manual.

Component 2: Given that the raters in our study had varying levels of prior education and experience working across the entire age range of our study population (i.e., infants through school age) we assembled professional development resources (e.g., state early learning guidelines courses, courses focused on evidence-based practice) that allowed our raters to independently build foundational knowledge across age ranges, with a particular focus on content related to infants, toddlers, and preschool aged children. Trainees also independently reviewed the *Texas Core Competencies for Practitioners and Administrators* paired with an online overview course which was intended to provide common understandings and expectations of the child care setting among raters.

Component 3: Raters engaged in multiple rounds of independent practice (i.e., scoring from authentic artifacts and videos of classroom interaction) followed by group feedback and discussion. During these discussion sessions, lead trainers took notes regarding key scoring challenges, clarifications, along with systematic documentation of examples and nonexamples to include in subsequent stages of training design and development. After 6 rounds of this mock scoring approach, raters participated in 3 site visits to practice data collection and scoring alongside master raters. Early sessions were characterized by side-by-side exposure to the master rater's process and thoughts related to TRS measures. Later observations were conducted alongside the master rater without discussion to allow for objective comparison of master rater and trainee documentation and scores. Raters were released to code independently

once agreement with the master rater reached a rating equal to or greater than .07 as well as sum of squares that was equal to or less than 2.0 across 5 consecutive assessments. After data entry and comparison of agreement, raters met with the master rater to discuss discrepancies in scores. Given that the measures vary by age and that raters have differing levels of knowledge and experience across age levels, we examined agreement by age group prior to releasing raters to score independently (i.e., some raters needed additional practice opportunities in order to demonstrate agreement in all age groups).

Our goal for training development in **phase 2** was to transition away from face-to-face didactic sessions and move toward self-paced web-based training content that would later feed into the online training and certification system. During onboarding for the study, trainees attended a kickoff meeting and were provided a training plan that included assignments to view a series of presentations with notes and embedded video content that had been adapted and sequenced based on implementation experiences and feedback from phase 1 (i.e., components 1 and 2 above). Training plans included embedded practice opportunities and master raters maintained a regular schedule of debrief meetings to discuss discrepancies in scores. In addition to these individualized meetings, all raters (i.e., those already released to collect data and trainees) attended a weekly one-hour check-in meeting to monitor agreement and resolve any concerns. These meetings were structured to yield information that could be used to build the online training and certification system (e.g., refinement of technical scoring manual language and examples) and to allow us to pilot a structure and process for running virtual small group feedback sessions (i.e., scalable format for state level monitoring and support) pre and post certification.

Finally, lessons learned from phase 2 training were incorporated into the design for the online training and certification program (i.e., online, self-paced training with virtual PLC and individualized support as needed) and routine performance monitoring procedures (i.e., quarterly monitoring routines supported by virtual PLCs). Additional details about this deliverable can be found in Appendix 9.

Establishing Inter-Rater Agreement

During the training phase, the average sum of squares of differences (AVE_SOS; see 2.2 Analysis Plan) for each item was calculated. We adopted a cut-point of 1.7 to determine whether raters can be released to assess independently. The cutpoint was determined by comparing different cutpoint values until a threshold was reached that aligned with a minimum G-coefficient of .7. Raters were required to meet the cutpoint across 10 consecutive observations (i.e., some raters required more than 10 observations to meet cutpoint). Of the 16 raters that participated in training, three raters resigned or were reassigned prior to data collection and three raters were not released for independent classroom (CARF) assessment.

Provider Recruiting Procedures

Data Sources for Participant Selection

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.

Using data from the US Census Bureau, we categorized communities (i.e., zip codes) into high, medium, and low SES groups using the percentage of families with children under five years of age whose income in the last 12 months was below the poverty level indicator. The range of these percentages for each SES group is shown below:

Study SES Percentage Range:

- High 0.0 – 7.1
- Medium 7.2 - 25.7
- Low 25.8 - 100.0

This aligns closely with the statewide SES Range:

- High 0.0 - 3.8
- Medium 3.9 - 27.1
- Low 27.2 - 100

SES classification allowed the team to strive for balance during recruitment, which increases confidence that the data captures potential variation in quality associated with SES and that the findings can be applied to a diverse set of providers.

Steps in Recruitment Process

Based on the data sources described above, a list of approximately 2,900 sites was included in the study database. A postcard describing the study was mailed to providers. Each member of the recruiting team received a list of about 350 schools to contact by phone. Each school was prescreened to determine if the site qualified for the study (i.e., site contained classrooms across the four age groups) and if they were interested in participating in the study. A site visit to discuss the study was provided upon request. If interested, recruiters would email a flyer with details about the study as well as director/teacher commitment letters for the entire site to complete. Recruiters would also update the site's information in the study database, including any accreditations, current participation of Texas Rising Star or Texas School Ready, the site's number of classrooms, and the number of caregiving staff in each classroom. Recruiters would also request lesson plans, daily schedules, and a list of all staff members from the providers.

Inclusion and Exclusion Criteria

In order to participate, sites needed at least four classrooms, one per age group: infant, toddler, preschool, and school-age. 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.

Sites were excluded if any of the following conditions were met:

- Site was less than one year in operation
- Site did not have Infant or Toddler Classrooms
- License revoked/suspended in the previous five years
- Site was included in video samples used to support the development of the TRS Assessment Training and Certification Program (described in Appendix 9).

Recruitment results can be summarized as follows:

- Total number of sites contacted to reach full sample target: 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
- Not contacted: 200
- Total withdrew: 14
- Final study sample: 128 providers

Final Sample Characteristics

As indicated above, 128 providers participated in the study, 69 of which were TRS certified prior to or during the study period. To fully understand the characteristics of the final sample, we examined this population through two lenses: socio-economic status of the community and licensing history of the provider.

Classrooms by Socioeconomic Status (SES)

We categorized centers as serving low, medium, or high SES communities using the same method described in the recruitment section. Because SES status is often highly correlated to achievement gaps of children (Cheng & Peterson, 2018; Duncan & Murnane, 2011; Duncan, Morris, & Rodrigues, 2011; Hanushek, Peterson, Talpey, Woessmann, 2019; Heckman & Karapakula, 2019; Magnuson & Waldfogel, 2008), a goal of the study was to recruit a balance of SES levels in order to examine the extent to which ratings varied among the three groups. The following table presents the number of classrooms per age range across the three SES levels.

Classrooms by Socio-Economic Status

Values	Low	Medium	High	Total
Sum of Infant (0-17 mths)	58	72	59	189
Sum of Toddler (18-35 mths)	69	96	82	247
Sum of Preschool (3-5 yrs)	68	113	99	280
Sum of School Age (5-12 yrs)	44	62	42	148
Total	239	343	282	864

Licensing History of Participating Centers

Texas Health and Human Services is the child care licensing (CCL) and regulatory agency for the state of Texas. Providers must demonstrate consistent compliance with minimum state CCL requirements to participate in TRS. Providers placed on corrective or adverse action by CCL are automatically found not to have demonstrated consistent compliance with minimum licensing standards and, therefore, are not eligible to participate in the TRS program. A child care facility is not eligible to apply for TRS certification if, during the most recent 12-month CCL history, the provider had:

- any critical licensing deficiencies, as listed in the TRS guidelines;
- five or more high or medium-high licensing deficiencies, as listed in the TRS guidelines; or
- 10 or more total licensing deficiencies of any type.

For certified providers, five high to medium-high deficiencies or a single critical deficiency results in the loss of a star-rating (e.g., reduced from 4-star to 3-star) or the loss of certification for 2-star-rated providers. Moreover, a TRS certified provider will be put on TRS probation when 10–14 total CCL deficiencies are cited within a 12-month period, and 15 or more deficiencies result in a loss of TRS certification.

Using data from the CCL, we examined licensing history for sites participating in the study to examine the extent to which patterns of deficiency relevant to TRS standards varied in our sample of providers. Most providers in our sample met the licensing thresholds for TRS eligibility. Among those that did not meet thresholds, we found:

- 20 providers exceeding the total number of deficiencies allowed in the last 12 months (10 or more)
- 23 exceeding TRS-selected critical deficiencies in the last 12 months
- 0 exceeding TRS-selected high/medium high deficiencies in the last 12 months

Information about the specific deficiencies cited (e.g., background check renewal) for participants in the study is in Appendix 1.

Among those providers that exceeded TRS-selected thresholds, most were corrected within several weeks, with the following percentages by type:

- Total deficiencies in 12 months; 96% corrected within 4 weeks
- TRS-selected critical deficiencies; 91% corrected within 7 weeks
- TRS-selected high/medium high deficiencies; 91% corrected within 4 weeks

Finally, we looked at the extent to which child care licensing deficiencies related to TRS category scores, and found small to moderate significant correlations between category 5 scores (Parent Engagement) and total number of deficiencies in a 12-month period (r= -0.47, p<.05), and the total number of high/medium high deficiencies in a 12-month period (r= -0.26, p<.05). This means that providers with fewer licensing deficiencies were also more likely to have formal family-related policies and procedures, and communication routines. We did not find significant correlations with other TRS category scores.

Assessment Procedures

Assessment Scheduling

Assessments were scheduled at least two weeks to one month in advance. Providers that were recruited and successfully completed the consenting process were contacted via phone to schedule a time for observation. Providers received a follow-up email containing pertinent information including the confirmed scheduled date and what to expect during assessment visits. A reminder call was also made the day before the scheduled assessment to confirm class information and remind the provider of the structure for the day of observation. Scheduling was based on the type of assessment (consistency or stability) as well as the number of classrooms at the facility. This information also helped to determine how many assessors were needed to complete classroom observations at a facility. For example, if the facility had 6 classes, there were 2 assessors assigned to complete classroom observations. Each assessor was able to complete a maximum number of 3 classroom observations per day at one facility. There was also one assessor designated to collect and review documents for the facility as well as complete scoring for those facility measures related to category 1, 4, and 5. The results of the assessment visits were not shared with the research study sites.

Onsite and Offsite Assessment Procedures

Please refer to Appendix 6 to review the sample forms used in the study (Facility Assessment Record Form (FARF), Classroom Assessment Record Form (CARF), Note-taking Form, and Director and Caregiver Worksheets).

Document Review

To score items requiring document review (category 1,3,4, and 5), assessors:

- distributed a document checklist for required and points-based measures to directors

immediately after recruitment.
- provided reminders of required documents during assessment scheduling call.
- provided reminders the day prior to assessment of the documents needed for review.
- discussed requirements during a walkthrough the day of assessment.
- requested specific missing documents during the onsite review process.

Classroom Observation

To score items during the classroom observation (category 2, 3-IFAL), assessors:

- completed a teacher interview with four questions before the observation.
- took notes on the note-taking handout during the one-hour observation.
- reviewed the classroom environment.
- reviewed the outdoor environment.
- completed a teacher interview with two to seven questions after the observation. Assessors only asked questions about what was not observed during the one-hour observation.
- took notes during 15 minutes of mealtime or snack time if it occurred outside the one hour observation.
- took notes during 15 minutes of outdoor time if it occurred outside the one-hour observation.

Assessors immediately scored category 2 and category 3-IFAL once the one-hour observation ended. The assessors determined the final item score by reviewing the notes from the note-taking handout, referencing the Technical Scoring Manual (TSM), and the coding updates spreadsheet. Assessors were then allowed to begin their next classroom observation.

Lesson Plans and Daily Schedules

The items for lesson plans and daily schedules were scored outside the one hour classroom observation (category 3). Assessors:

- reviewed lesson plans to determine whether there were four consecutive weeks of lesson plans and whether they had objectives.
- if there were four consecutive weeks of lesson plans and they had objectives, then the assessors completed a lesson plan table to help track the number of activities within a domain of the state guidelines for prekindergarten or infant, toddler, and three-year-olds.
- reviewed the daily schedule to determine the balance of caregiver-led and child-led activities and the amount of physical activity the children experienced during the day.
- scored each item by reviewing the lesson plan table, TSM, and the Coding Updates spreadsheet.

Facility Review

To score items during the facility review (category 1, 4, and 5), assessors:

- took notes on the facility environment.
- worked with the director or assigned staff to pull the requested documents for review.
- completed a director interview with 5 questions.
- reviewed the director and caregiver TECPDS reports to complete the director worksheet and

caregiver worksheet.
- reviewed facility and staff documents that could not be scored using the TECPDS reports.
- scored each item by reviewing notes that were taken, the Director Worksheet, the Caregiver Worksheets, and referencing the Technical Scoring Manual, and the Coding Updates spreadsheet.

Data Verification

To ensure that each item was completed, assessors reviewed their own forms first to make sure that each item was scored and all data completed and then turned in their data to be verified. Once the form was turned in, a verifier:

- reviewed the form a second time to make sure that each item was complete.
- tracked missing items to determine common errors that needed to be addressed with the individual assessor or assessment team.
- returned the form back to the original assessor if an item was missing to correct the form.
- verified the item was corrected and then turned in the data.

Analysis Plan

Several quantitative analytical approaches were applied in the reliability study to appropriately study the reliability of the TRS assessment. In this section, we documented analytical approaches used for each of the key questions under aims.

Aim 1a is to determine within and across category functioning of TRS measures. The following analyses were conducted to address research questions of interest. First, we investigated descriptive statistics including mean, standard deviation, and skewness, as well as histograms of item scores/values. Floor and ceiling effects were identified by histograms of frequency items that show an extremely high percent of value of 0 (floor) or "equal or larger than 6" (ceiling). Floor and ceiling effects were also confirmed by the values of skewness - "equal or larger than 1.0" (right skewed - floor effect); "equal or smaller than -1.0" (left skewed - ceiling effect). In addition, items with histograms showing an extremely high percent of the item value not applicable N/A were considered to be items that were frequently excluded from scoring.

Second, different types of analyses were conducted to investigate which items are not contributing well/meaningfully to the assessment (subscale/overall) in current form. Those analyses included:

- correlation between item score and total score (item-total correlation),
- internal consistency (Cronbach's alpha),
- Cronbach's alpha if item deleted,
- generalizability coefficient, and
- factor analysis.

Item-Total Correlation

The item-total correlation is the correlation coefficient between the individual item score (e.g., P_LFS_01 from category 2) and the overall category score (e.g., category 2 total score). In the TRS assessment, items from the same category are designed to rate constructs (e.g., language facilitation and support, play-based interactions and guidance) that are relevant to a broader construct (e.g., category 2: Caregiver-child interaction). Therefore, the item scores coming from the same category are expected to be reasonably and positively correlated with the overall category score. We adopted a item-total correlation of +0.2 as a cutoff value. Items with item-total correlations equal or less than +0.2 can be viewed as weak items (e.g., item did not belong to the corresponding category, item did not measure what it was supposed to measure, item description was not clear).

Internal Consistency

The internal consistency (Cronbach's alpha) is a measure of score reliability based on the correlations between item scores within the same category. Cronbach's alpha indicates the extent to which a set of items are closely related as a group. For these specific research questions, the Cronbach's alpha was computed for each category and then was used to compare with the Cronbach's alpha if item deleted (see below).

Cronbach's alpha If Item Deleted

The Cronbach's alpha if item deleted (Cronbach's alpha-ID) is the value of Cronbach's alpha after the targeted item is removed from the category. Cronbach's alpha-ID was first computed for each of items and then compared with the original Cronbach's alpha value by category. When an item has a Cronbach's alpha-ID larger than the original Cronbach's alpha value (i.e., removing the item leads to higher internal consistency), we considered removing this item.

Generalizability Coefficient

The generalizability coefficient (G-coefficient) is computed based on the generalizability (g) theory (Marcoulides, 2000; Shavelson, Webb, & Rowley, 1989). Simply speaking, g theory estimates the variation in scores due to each person (e.g., teacher), each facet (e.g., items, assessors, occasions), and their combinations (interactions). G-coefficient for an absolute decision was computed for the TRS assessment (see the formula in Marcoulides, 2000). We adopted the G-coefficient as a measure of score reliability.

Factor Analysis

Factor analysis is a statistical method used to explore or confirm the number of underlying constructs and examine the extent to which the items are designed to measure the same construct. We conducted confirmatory factor analysis (CFA) in the framework of structural equation modeling. A two-step procedure was employed: First, for each of the categories, we specified a

model assuming one underlying construct exists (i.e., one-factor model). If the one-factor model can fit the data well (e.g., category 2 and 3), we then concluded items from the same category were measuring the same general construct. Model fit indices and corresponding cutoff values (root mean square error of approximation [RMSEA<0.06], comparative fit index [CFI>0.95], Tucker-Lewis index [TLI>0.95]) were applied to inform the goodness-of-fit (Hu & Bentler, 1999). If not (e.g., category 4), an exploratory factor analysis (EFA) was applied to explore the number of underlying factors. A new CFA assuming multiple underlying constructs was applied again to confirm the results of EFA. During the model specification, we allowed some residuals of items to be correlated to reflect the practical reality for TRS assessment. The statistical package Mplus was used for the analysis.

Note aforementioned analyses were iteratively conducted to test whether the reliability of ratings, within and across categories, can be improved through removal of items or by using alternate scoring criteria.

Aim 1b is to determine inter-rater agreement and reliability within and across TRS categories. Multiple statistical indicators were used to evaluate the inter-rater agreement and reliability. For example, the percent agreement for two raters was computed by dividing the number of ratings in agreement by the total number of ratings, and then converting the result to a percentage. A higher value of percent agreement suggested better inter-rater agreement.

We also computed the average sum of squares of differences between two raters (AVE_SOS). To get the AVE_SOS, first, we squared the deviation score between two raters for each of the items and then sum them up (i.e., the sum of squares of differences). Note although the deviation score can indicate the degree of disagreement, squaring the deviation score is in a sense that a larger disagreement in rating scores should be more weighted when we evaluate the inter-rater agreement. Second, we divide the sum of squares of differences by number of items to receive the AVE_SOS. A lower value of the AVE_SOS suggested better inter-rater agreement. Last, we computed the G-coefficient to evaluate inter-rater reliability between two raters and between multiple raters.

Aim 1c is to examine the stability of star ratings over time. Changes in star ratings between assessments were first evaluated by crosstabs that show the relationship between star ratings at different timepoints. In addition, for classroom data with two timepoints, we have used the SAS Proc Mixed procedure to account for the dependency of classrooms that are from the same school and the results suggested the dependency was trivial. As a result, the paired T-test was applied to compare the mean differences of category scores over time. A statistically non-significant result suggested the stability was held. For classroom data with three timepoints, the growth modeling approach was used to test the growth of category scores overtime using the SAS Proc Mixed procedure. A statistically non-significant growth (i.e., the slope parameter of the growth model) suggested the stability was held.

Aim 2 is to examine for indicators of validity of the TRS assessment across categories and with other measures of quality and outcomes. To address this aim, quantitative analyses included descriptive statistics to compare the distribution of TRS scores (by category) among regions with different socio-economic status, the Pearson correlation coefficients to evaluate measures across categories relate to each other as well as the strength of the correlation between the TRS

assessment and other measures of quality and outcomes (e.g., national accreditation status, child outcomes).

Section 3

Results

This section begins with descriptions of findings within each category that resulted from an item-level screening process. The screening process aimed to identify poorly functioning items that could potentially be revised or removed. This is followed by a cross-category analysis section, which examines whether item-level changes improve reliability of the instrument. The third section includes findings from an initial exploration of external validity. Finally, the section ends with limitations of the study. For readability, we often reference items by their alphanumeric codes. We encourage readers to refer to the Facility Assessment Record Form and Classroom Assessment Record Form in Appendix 6 for the full item text.

Key Definitions for Analysis and Recommendations:

- ★ **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.

Category-Level Item Screening

We began by identifying items with low score variation, frequent exclusion (scored N/A), and low correlations with total scores at the category level. We looked at the percentage of providers that met criteria for met/not met items and the normality of score distribution for points-based items. We considered multiple aspects of item functioning to inform recommendations for removal or adjustment. Removal or adjustment of items is intended to strengthen the reliability of scores (e.g., remove items with low contribution to quality scores) and to reduce the scoring

burden on raters (e.g., prioritize removal of items with low contribution to scores that are time consuming to assess).

Normality of Score Distributions of Points-Based Items

We looked at the distribution of scores to determine skewness of the ratings. We considered skewness values "equal or below -1.0 (left skewed)" or "equal or above 1.0 (right skewed)" as highly skewed distributions. Scores with left skewed distributions indicate ceiling effects (skewed to maximum scores), while scores with right skewed distributions denote floor effects (skewed to minimum scores). In some cases, the distribution was improved using an alternate scoring. Please refer to the Item-Level Descriptives for Points-Based and Met/Not Met Items table in Appendix 2.

Cronbach's alpha Cutpoints

Cronbach's alpha is a measure of internal consistency. Simply speaking, Cronbach's alpha shows the extent to which a set of items are closely related (inter-item correlation) as a group. When a set of items has a low Cronbach's alpha value, it is likely that some items are measuring something else, item score has close-to-zero variability, or large measurement error is introduced. Another determinant of Cronbach's alpha value is the number of items—a smaller number of items leads to a lower Cronbach's alpha value.

- Below .6 = unacceptable
- .6 to .69 = borderline acceptable
- .7 to .79 = acceptable
- .8 to .89 = good
- .9 and above = excellent

In research settings, lower levels of Cronbach's alpha are considered acceptable (i.e., acceptable range); however, in a policy context (i.e., where there are funding implications), it is advisable to look for good to excellent levels (e.g., greater than .9) of internal consistency to increase confidence in the ratings system.

In the sections that follow, we present the category-level results that inform recommendations for item-level removal or revision. We first present key findings related to met/not items (i.e., 2-Star requirements followed by points-based items (i.e., contribute to 3 and 4-Star certification). Full descriptive information for TRS items can be found in the Item-Level Descriptives for Points-Based and Met/Not Met Items table in Appendix 2. It is important to note, that in our sample many providers would not have met the 2-star requirements for participation in TRS, and therefore would not have been assessed on points-based items under routine TRS practice. Given that the purpose of this study was to learn about instrument functioning under the current quality standards, rather than the program's policies, we collected full assessments on all participating providers.

Category 1

Overview

Category 1 includes measures 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 measures. 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 although 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 below.

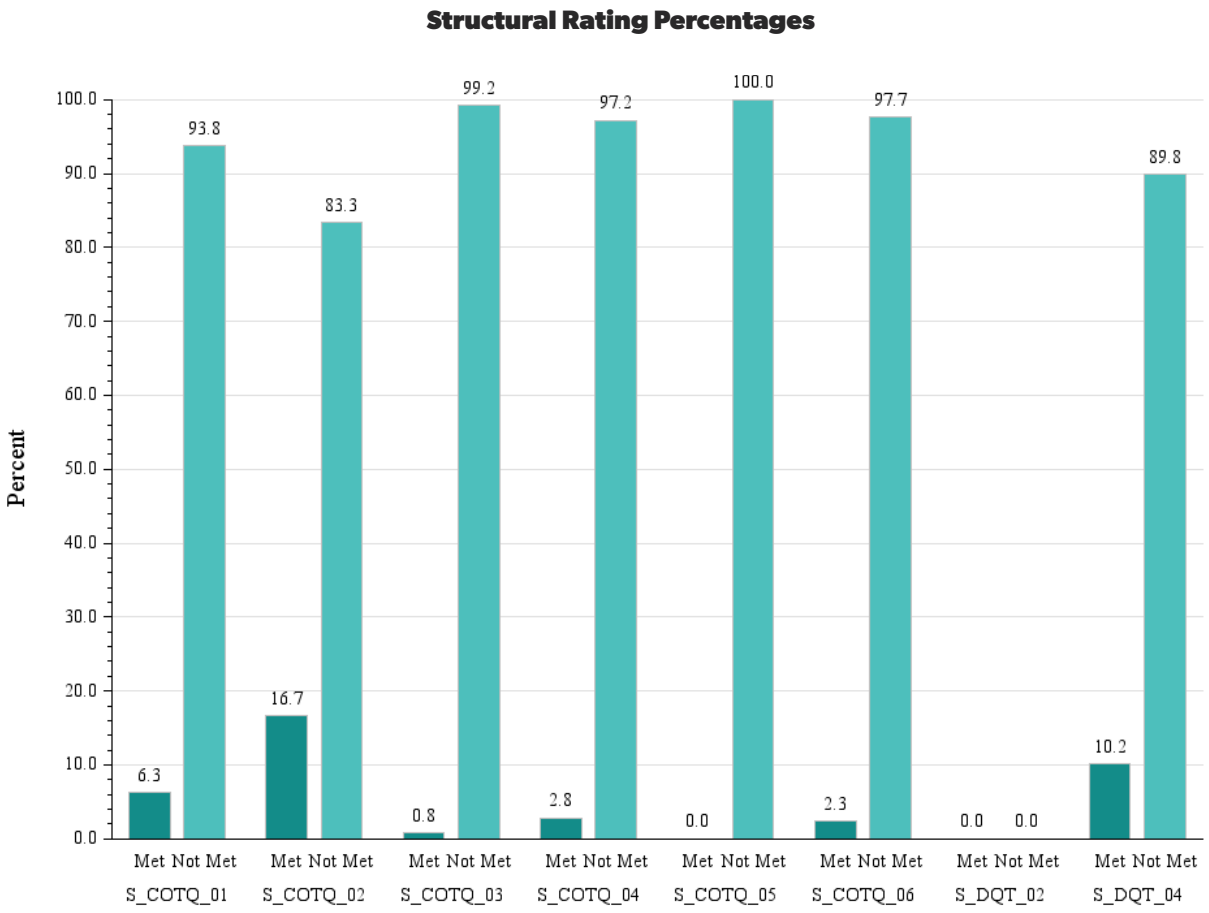
Subcategory	Number of Points-Based Items	Number of Met/Not Met Items
Director Qualifications	3	2
Caregiver Qualifications	2	6

Scoring for many of the items within this category happens through document review. To facilitate scoring of items requiring document review, assessors:

- distributed a document checklist for required and points-based measures to directors immediately after recruitment
- provided reminders of required documents during assessment scheduling call
- provided reminders the day prior to assessment of the documents needed for review
- discussed requirements during a walkthrough the day of assessment
- requested specific missing documents during the onsite review process

Met/Not Met Measures

In the study sample, no centers met criteria for all of the items required for a 2-star level rating. No individual items were scored as met by more than 17% of providers. The following chart indicates the percentage of centers who met and did not meet each item (excluding those that marked N/A for the item).



Items Frequently Excluded

Additionally, several of the non/not-met indicators were frequently excluded from scoring in our sample which means these items could not be applied equally across providers. This suggests these items are not consistently contributing information to provider scores as currently written. It may be that many of these items are addressed during TRS pre-assessment technical assistance activities. Items scored Not Applicable (N/A):

- S_COTQ_02 volunteer and substitute caregiver orientation, 86%
- S-COTQ-04 full-time caregiver staff training-school age, 45%* (N/A allowed if caregiver employed for less than 90 days)
- S-COTQ-05 part-time caregiver staff training- school age, 61%* (N/A allowed if caregiver employed for less than 90 days)
- S_DQT_02 TRS director certification course, 100%.

Points-Based Measures Analysis

Director-Level Education

The table below shows the highest level of education obtained by directors in the study sample.

Director Qualification and Credential

Highest Education Level Achieved Scores: 1=High school/GED, 2=Associates degree, 3=Bachelor's degree, 4=Master's degree, 5=Doctorate. Note: * information was not available in files or upon request.

Score	Frequency	Percent	Cumulative Frequency	Cumulative Percent
*	53	41.41	53	41.41
1	45	35.16	98	76.56
2	8	6.25	106	82.81
3	13	10.16	119	92.97
4	9	7.03	128	100.00

For item P_DEQT_01, the indicator-level scores show that only 40% of directors meet at least one criteria captured by this item. The table below shows the percentage of administrators in our sample that met each indicator.

Study Consistency Observations from Director Worksheets: Director Education (N=128)

Criteria	% Met Criterion
Meeting at least one of the Criterion	39.2
Valid child care administrator's credential	23.2
Valid Child Development Credential (CDA), or Child Care Professional (CCP) Credential with 6 college credit hrs in business management	1.6
9 college credit hrs in ECE and 9 credit hrs in business management	0.0
60 college credit hrs with 9 college credit hrs in child development and 6 college credit hrs in business management	0.0
A child care administrator's certificate from a community college with at least 15 college credit hrs in child development and 3 college credit hrs in business management	0.8
Over 4 years, up to 8 years as a director in a TRS or TRS- recognized nationally accredited provider	0.8

Criteria	% Met Criterion
AA/AAS in ECE or closely related field with 12 college credits in ECE and 6 credit hrs in business management	0.0
At least a BA/BS with 12 college credit hrs in ECE and 6 credit hrs in business management	5.6
Over 8 years as a director in a TRS or currently recognized nationally accredited provider	1.6
Non- expiring director's certificate from DFPS	5.6

Director-Level of Early Childhood Experience

The distribution of scores for director experience (P_DEQT_04) suggests substantial loss of potentially meaningful information is occurring under current scoring criteria. For example, within the TRS assessment a director needs four years of experience to receive a score of 4, and in the study sample directors have an average of 11 years of early childhood experience (*SD*=9.4). TWC can consider adjusting the scoring criteria to allow for a more complete picture of variation in ECE experience. The current scoring criteria along with recommendations for a new range are below.

PDEQT_04: Current Scoring

Years of experience in ECE programs	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-1	15	12.1	15	12.1
2	7	5.65	22	17.74
3	7	5.65	29	23.39
4 or more	95	76.61	124	100

Frequency Missing = 4

PDEQT_04: New Recommended Range

Years of experience in ECE programs	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-1	15	12.1	15	12.1
2-5	27	21.77	42	33.87
6-10	33	26.61	75	60.48
10 or more	49	39.52	124	100

Years of experience in ECE programs	Frequency	Percent	Cumulative Frequency	Cumulative Percent

Frequency Missing = 4

Director-Level Training

The following table shows the distribution of scores for director annual training (P-DEQT-06).

- 83% of directors received a score of 0 on the annual training item (i.e., had less than 36 hours of annual training).
- 82% had no information regarding program administrator training
- 95% had no information regarding infant and toddler state guidelines training
- 97% had no information regarding prekindergarten state guidelines training

The total score for this item is almost entirely attributable to total training hours and program administration-specific hours. In the Recommendations section, we describe recommendations for incorporating guidelines training into TRS QI plans to better emphasize the importance of this TRS standard and ensure providers account for it in their training plans (Recommendation 6).

Study Sample Scores for PDEQT_06

Score and Criteria	Number of Directors	Percent
N/A = New hire or initial applicant	5	3.91%
0 = None	106	82.81%
1 = Director has 36 hrs, 6 hrs in program admin,management & supervision	14	10.94%
2 = Director has 36 hrs,6 hrs in program admin & 3hrs in Infant/Toddler or Pre-K guideline	1	0.78%
3 = Director has 36 hrs, 6 hrs in program admin & 3 hrs in Infant/Toddler & 3 hrs in Pre-K guideline	2	1.56%

Teacher-Level Education

The majority of caregivers in the study sample had a high school diploma/GED without a CDA credential (69%) as the highest level of education achieved. Approximately 16% of caregivers earned a CDA, and 3% were working towards a college degree or have child development related college credit hours. Only 9% of the sample had a bachelor's or master's degree. The following table details the percentages of caregivers who met each of the seven possible criteria,

which combine the indicators above along with additional criteria. Only 20% of caregivers in the study sample met one of the criteria for Caregiver Qualifications and Training (P-CQT-01).

Consistency Observations from Caregiver Worksheets: Caregiver Qualifications and Training (N=1375)

Criteria to score P_CQT_01	% Met Criterion
A: Have CDA credential	7.3%
B: Have CCP credential	0.0%
C: Working towards an Associate's or Bachelor's OR have completed 12 college credit hrs in ChildDev/Earlychildhd edu AND 2 yrs of full time experience as caregiver with children in licensed/ registered facitlity	3.0%
D: Have 2 yrs of full time experience as caregiver with children in licensed/ registered facility while working toward CDA or CCP credential	1.3%
E: Have 150 training clock hours within the last 5 years in ChildDev/ Earlychildhd edu and 2 yrs of full time experiences as caregiver with children in licensed/registered facility	2.3%
F. Have Associates, Bachelor or Master	8.9%
G: Ten years of full time paid experiences as a caregiver in a TRS or TRS-recognized nationally accredited center	0.2%
Meet at least one of the criterion	19.6%

Teacher-Level Training

Annual clock hour training varied widely among caregivers, ranging from 0-150 hours, with an average of 12 hours per year. Scoring for caregiver training plan alignment (P-CQT-03) requires assessors to determine the extent of alignment between the core competency areas associated with specific trainings found in individual caregiver training plans, and the age group supervised by each caregiver. Once caregiver level alignment has been determined, assessors sum the number of aligned training topics for all staff, divide by the number of training topics, and finally multiply by 100 to determine facility-level training alignment. Providers with more than 80% alignment in training topics receive a score of 3. In the study sample, 88% of providers scored 3 for this item. Given that this item takes an extensive amount of time to score and fails to differentiate quality among providers, we recommend removal.

Teacher-Level Experience

In general, the experience level of caregivers in our sample was low, with 75% of caregivers having worked in a licensed or registered facility for less than 5 years. Among caregivers, 38% had less than one year of experience, only 15% had one year, followed by 10% with 2 years, 8% with 3 years, and 5% with 4 years.

Points-Based Items Distributions

Items in category 1 were not evaluated using measures of internal consistency 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). The distribution of scores indicates multiple non-normal distributions. As currently scored, many of these items do not appear to contribute information that meaningfully differentiates qualifications and training.

CATEGORY 1 HIGHLIGHTS

- **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), including:**

  - 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. The study team developed worksheets that better facilitate scoring of the items, which improved the thoroughness and accuracy of review. When TECPDS was used to facilitate scoring, time estimates dropped to 10-15 minutes.

- **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.

Please see Recommendations for Item Revision or Removal in Appendix 4 for more details.

Category 2

Overview

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

Category 2 Number of Items by Age Group

Sub-Category	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

Comparison of Current and Alternate Group Size Ratio Scoring

We examined differences in scores when using enrollment data (i.e., current scoring criteria) versus staff and children present during the observation period. The latter calculation resulted in greater variation in scores and showed stronger correlations with caregiving behavior. The distribution of scores for both scoring approaches is shown below. For a breakdown of ratio by socioeconomic status within the sample, please see page 55.

Enrollment vs Present Group Size Ratio

Data Collection Method	Score 0	Score 1	Score 2	Score 3
Enrollment Information Review	23.1%	19.1%	22.6%	35.2%
Present during Assessment	8.9%	12.6%	21.7%	56.7%

Item-Level Screening for Remainder of Category 2

Ratings Distribution (CARF)

One approach for evaluating items is to consider the extent to which scores are differentiating quality among providers (i.e., key goal for QRIS). Specifically, we examined for floor or ceiling effects that suggest a substantial portion of the providers in the sample are not distinguishable from one another. Based on scoring patterns observed in field data, we identified a specific type of item that seemed susceptible to this problem (frequency-based) and developed alternate scoring criteria to test alongside the current scoring criteria. The table below shows the thresholds set in the study to categorize the normality of scoring distributions across all items and the severity of the ceiling effect for frequency-based items.

Distribution Thresholds

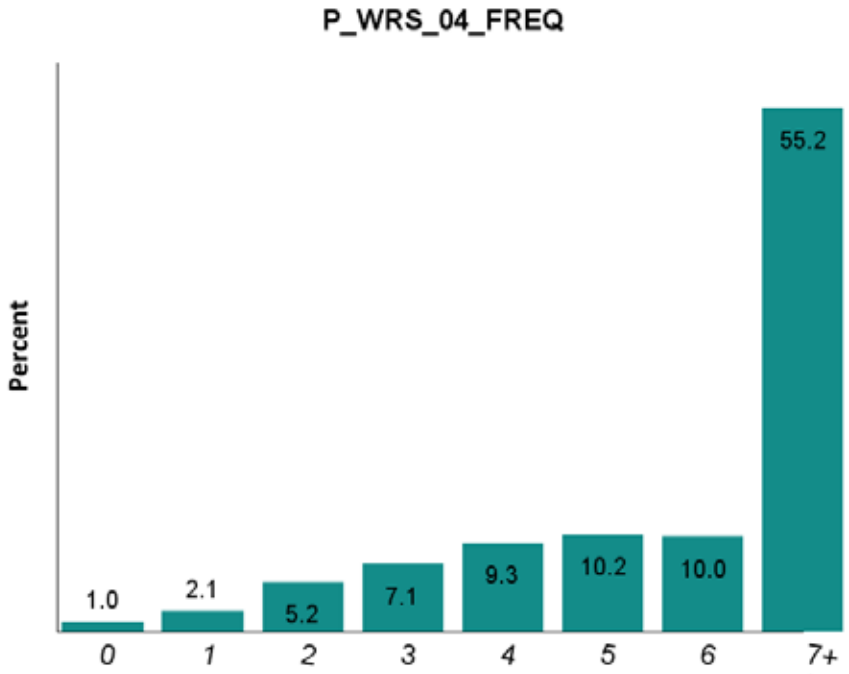
Distribution	<30% at frequency ceiling (6 or more)	>30% at frequency ceiling (6 or more)	Frequency not collected
Normal	Acceptable	Inconclusive	Acceptable
Skewed	Inconclusive	Not acceptable	Not acceptable
No distribution	--	--	Not acceptable

Floor and ceiling effects (i.e., highly skewed items) can be interpreted in multiple ways, and may not always indicate the need for item change. For example, a floor effect (i.e., all providers are unlikely to perform well on an item) might not raise measurement concerns if the criteria/ quality for the item is well supported by theory or external evidence (e.g., positive language input is highly desired based on extensive research but is not yet prevalent in the population). Conversely, a floor effect connected to item content not well supported by theory or evidence

may be viewed as punitive if the target behavior is not present in the provider sample yet prevents providers from receiving higher ratings. Ceiling effects impact assessment systems in a different way. If all providers in a sample receive the highest rating on an item, the system does not fully differentiate quality in that area of practice. This may be acceptable if the system is confident that the highest score possible is associated with the outcomes of interest (e.g., high score represents an important threshold of quality that relates to a positive child outcome). However, if the ceiling is set too low scores may not be able to predict outcomes (e.g., unable to validate system's impact on children), assessment-linked quality improvement efforts may be less targeted to actual needs, and reimbursement may be less differentiated than intended.

Ceiling Effects Associated with Frequency Maximums

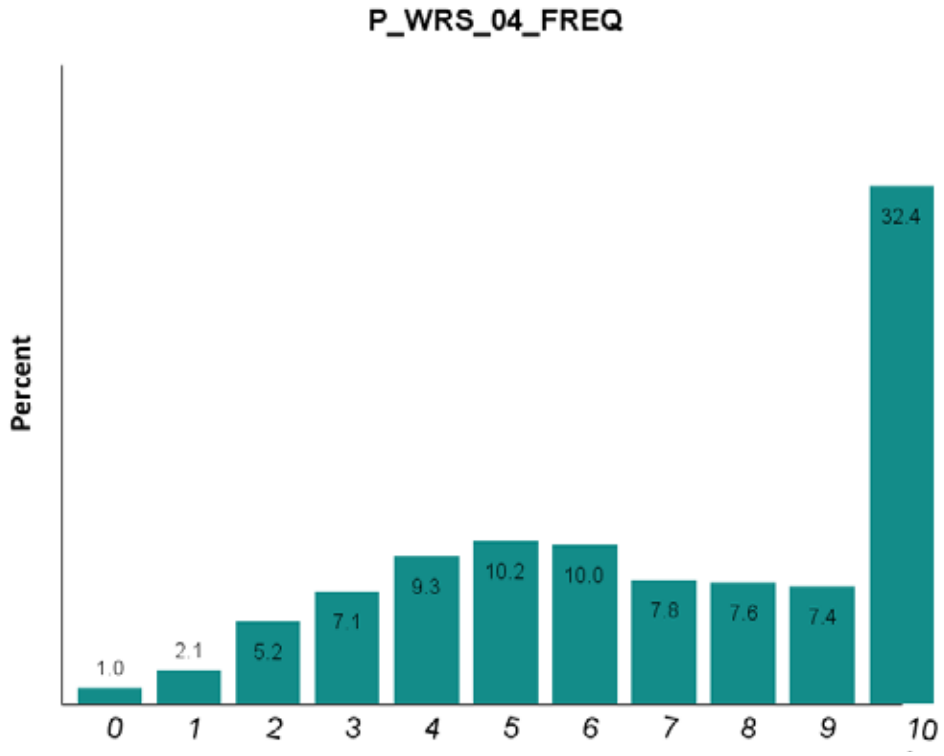
Frequency refers to the number of times the behavior is seen during the one-hour classroom observation period used to score category 2. In the current scoring criteria, a subset of items within category 2 rely on frequency counts to determine the rating. The current scoring criteria assigns points to specific frequency ranges (e.g., 0 points = zero to one instances, 1 point = two to three instances, 2 points = four to five instances, 3 points = six or more). One method used to evaluate item-level performance was to determine if the current minimum or maximum score allows raters to fully quantify the caregiver's use of the key behaviors across the one-hour observation period. We found that several items are experiencing a ceiling effect, in which a high percentage (defined as 30%) of the sample is receiving the maximum score. This indicates that the frequency scoring can be adjusted to achieve a more equal distribution of scores that reflects a greater range of caregiving behavior. We identified six items in category 2 that appear to have ceiling effects. For example, the following chart shows ceiling effects for WRS 04.



As the chart above demonstrates, more than 65% of the sample would receive a score of 3 (highest score possible). For 55% of the sample (those with 7 or more instances of the behavior), performance is indistinguishable among the highest performing providers in the sample.

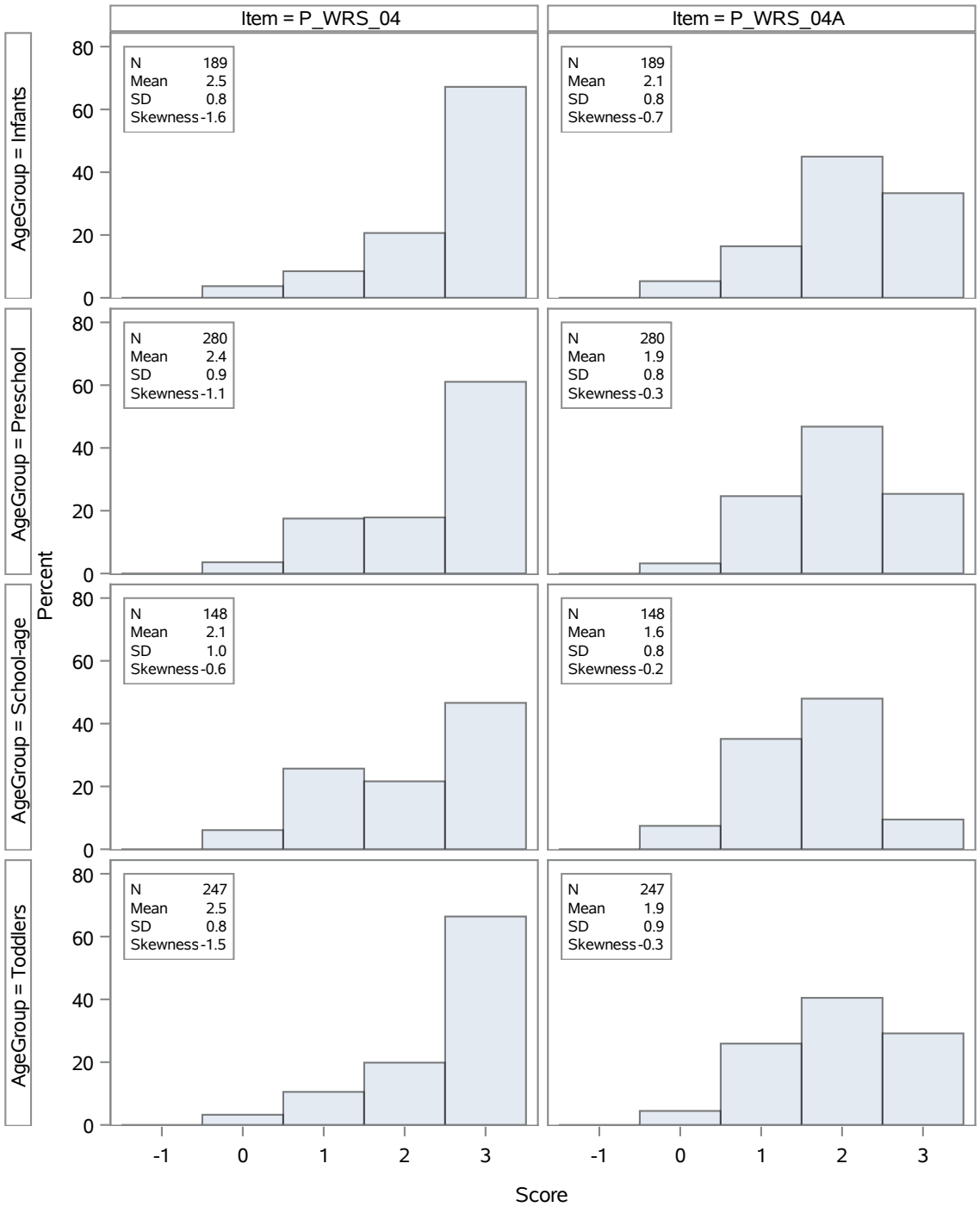
WRS 04 Alternate Scoring: Increased Frequency Maximums

For items that did not perform well with a threshold of six instances, we explored capturing as many as 10 instances to determine if the new frequency limit more fully represented our sample. The chart below shows the adjustment of providers receiving the highest score when we increased the frequency limit to 10 instances of the target behavior. As shown in the following chart, this adjustment allowed us to detect more variation in sample characteristics and reduce the percentage of providers for whom performance is indistinguishable (32%). However, this remains a substantial percentage of providers whose performance is not distinguished (i.e., does not eliminate ceiling effect).



This pattern suggests that the frequency limit would need to be raised substantially higher in order to capture a full range of typical behavior. The challenge for raters is that scoring frequency based items can be cognitively taxing. Given the substantial number of measures that raters must attend to during the observation period, it is likely that reliability would suffer if these frequency limits were increased. We tested an alternate scoring approach for these frequency measures to see if we could reduce the burden on raters, and improve item functioning (e.g., reduce ceiling effects). The alternate scoring method was designed to reduce the dependency on frequency-based scores, which require assessors to tally discrete events across many key

behaviors. The alternate scoring method reduces this cognitive load by allowing assessors to document more qualitative aspects of behavior that often serve as evidence across multiple items (i.e., one standardized note may include key information about language, warmth, and guidance items). Moreover, the alternate scoring reduces post-observation scoring time (i.e., reduced from approximately one hour to 20 minutes to finalize scores). The following chart provides an example of how the score distributions are improved using the alternate consistency-based scoring approach.



Category 2 Cronbach's alpha

Cronbach's alpha including all items at all ages was acceptable for both current and alternate scoring methods, indicating overall internal consistency of items within the category (values greater than .6 are considered acceptable; values above .90 are considered excellent). Given the implementation benefits described in the previous section and that Cronbach's alpha was slightly improved for the alternate scoring method, a shift to using the alternate scoring method is recommended. The Cronbach's alpha values for category 2 can be found in the table below.

Category 2 Internal Consistency

Age Group	Traditional	Alternate
Infants	0.90*	0.93*
Toddlers	0.91*	0.93*
Preschool	0.91*	0.93*
School-age	0.90*	0.92*

**Cronbach Alpha >.70*

Note: Includes P-SCR-01 and P-SCR-03 - Correlations with the total score were low for preschool (.17). However, Cronbach's alpha and factor analysis both support retention of this item.

CATEGORY 2 HIGHLIGHTS

- **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

Overview

Category 3 includes measures 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 measures.

Category 2 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

Item-Level Screening and Variability in Scores

Lesson plan items are scored based on criteria (i.e., number aligned lessons per week) applied to the following key components: 1) developmental domains (e.g., language, social and emotional), and at the preschool level specific academic skills areas (e.g., preschool literacy, math); 2) learning objectives linked to activities; and in some cases, 3) the developmental appropriateness of the documented approaches. We identified an initial concern that led the team to create

alternate scoring criteria for lesson plans and curriculum to examine alongside the current scoring method. In the current scoring approach alignments are taken as fact (i.e., accurate) if the provider/staff have documented (i.e., noted on the lesson plan) the learning objectives themselves. If this information is missing, raters are required to make this determination themselves and proceed with scoring. We wanted to assess the reliability and item properties within this area using a consistent scoring method in which assessors score these items based on their own training about alignments.

Unfortunately, scores for nearly every item, across age groups, show floor effects (i.e., most classes receive a score of 0). The only item with normally distributed scores represents physical activity and motor development (P-LPC-15), all ages. With the support of templates and mentoring it is likely that most providers would be able to generate lesson plan documentation in alignment with the current scoring criteria.

Scores within Instructional Formats and Approaches to Learning were more normally distributed. Two items, IFAL 03 (using routine and transition times for incidental learning) for infants and school age, and IFAL 05 (repeated exposure of new concept in different contexts) all ages had skewed distributions. Given the brevity of the observation period, it is not surprising that these items are observed less frequently.

For two items (P-SNRD-1, Consideration for children in a bilingual program and P-SNRD-2, Consideration for children with disabilities), floor effects resulted due to all caregivers receiving a score of 0. No variability also occurred when a high percentage of caregivers were excluded from rating with a "not applicable" score. These items are scored based on documentation of specific adaptation or support strategies across a week or month. Although we believe that observing these practices would provide more valuable quality information than documentation alone, we do not recommend adjusting the scoring criteria to allow for ratings based on observation with the TRS assessment. The range of behaviors associated with these skills varies greatly based on classroom composition, and given the relatively short observation period (i.e., one-hour) demonstrations of these skills will likely be rare, leading to inconsistent inclusion of the items across providers (i.e., many N/A will scored). Given that items are not performing well and there are too few of them to constitute a reliable measure of these sophisticated caregiving skills (see below), we recommend removal of the items as currently scored and provide recommendations for revision or including these critically important topics in TRS-supported quality improvement plans.

Category 3 Cronbach's alpha

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).

To determine if internal consistency could be improved, we removed items with low correlations to the total score or with limited variability in scores. Removal of the PSNRD items (planning for special needs and respecting diversity) led to marginal improvements in internal consistency for

infants, toddlers, and preschool, and more substantial improvements for school-age. However, internal consistency still falls in the unacceptable range for school-age. Internal consistency for infant items moves into the acceptable range (.72) by removing PSNRD items; however, this still does not meet the goal standard of .8 or above.

We also explored removal of additional items that had low correlations with the total score and found only marginal improvements to internal consistency that did not move Cronbach's alpha into a new acceptability range.

Given that removal of these items will not significantly impact the internal consistency of the category 3 scale, and given that these items have very limited variability and are often excluded (i.e., marked N/A, range 46-93% excluded), removal or revision of these items as currently written and scored is recommended.

The items within IFAL are focused on caregiving behavior and the distributions are more normal, suggesting that IFAL items may perform well within category 2. Correlations between IFAL items and category 2 scores are all highly significant and in the moderate to large range (r= .42 to .56, p<.01).

CATEGORY 3 HIGHLIGHTS

Category 3 is not functioning well in terms of internal consistency and distribution of scores. Substantial conceptual changes to category 3 are recommended to more meaningfully account for curriculum-related practices with TRS, as described below.

▶ 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 as currently written. 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 Recommendations section (Recommendation 1).

▶ 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

Overview

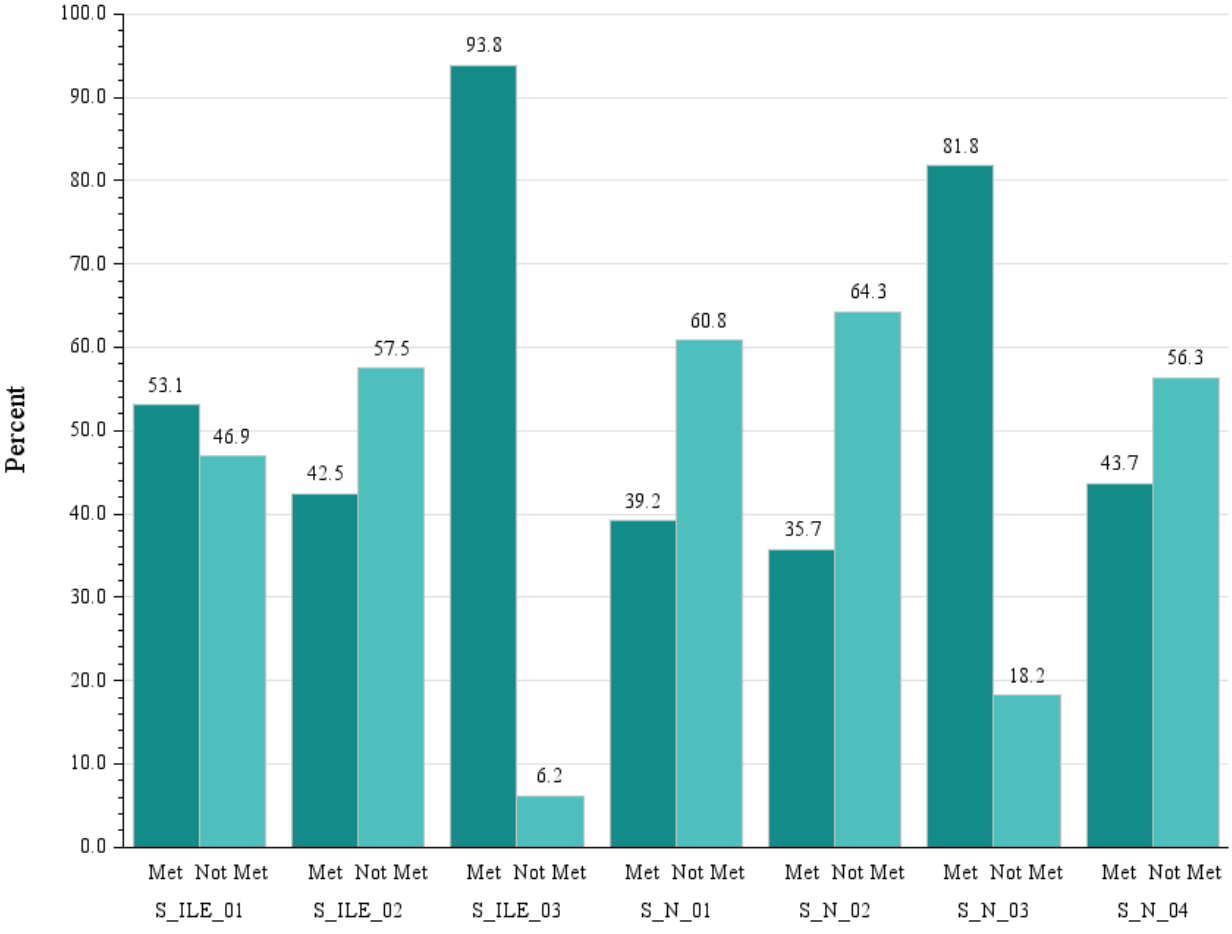
Category 4 includes measures related to nutrition policies and practices, as well as the equipment, materials, and arrangement of indoor and outdoor learning environments. The nutrition and indoor learning environments sub-categories include a combination of met/not met (required) measures and points-based measures. The outdoor learning environment sub-category is scored using points-based measures only.

Category 4 Number of Items by Age Group

Subcategory	Number of Items by Age Group				# of Met/ Not Met Items (at facility-level)
	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

The following chart shows the percentage of providers in our sample that scored "met" and "not met" across category 4 items, excluding providers with a score of N/A.

Structural Rating Percentages



In general, there was variation in provider scores across these items. Some exceptions included:

- S-ILE-03 which captures facilitation and completion of homework in school age classroom was scored as met by 94% of providers
- S-N-03 which captures menu planning policies and dietary review was met by 82% of providers

With few exceptions, item level distributions for points-based measures within category 4 were acceptable. Some notable concerns included:

- Item P-N-01 considers to 6 specific mealtime practices and scored at a 3 for more than 83% of providers
- Item P-N-03 was often excluded (38%) because the majority of children in the infant classrooms were receiving solid foods
- Item P-N-04 was often excluded (37%) because all children in the observed infant classroom were above 12 months of age
- P-OLE-01 which considers the extent to which the outdoor environment activities are linked to indoor learning was scored as 0 for 80% providers

Category 4 Cronbach's alpha

Category 4: Indoor Learning Environments for all ages is working well. Outdoor Learning Environments is well with the exception of the infant age group. Nutrition items are not functioning well on their own and they do not combine well with ILE or OLE constructs.

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).

We removed items that have low correlations to the total score in an attempt to reach internal consistency in the good range. After removing two infant nutrition items (PN03, "Infants are held and talked to in reassuring tones while bottle fed" and PN04, "Caregivers feed infants on the infant's cue...and stop feeding upon satiety") that have low correlations with the total and were often excluded (i.e., scored N/A 37% of the time*), infant internal consistency was improved to the acceptable range (.78). It is recommended that these items are removed for infant scoring. Moderate significant correlations for these items were found with warm and responsive behaviors captured in category 2, suggesting that these nutrition related concepts are closely related to caregiving behavior measured elsewhere (P-N-03 $r = .36$; P-N-04 $r = .38$, $p < .01$).

Removal of P-N-01, P-N-02, and OLE-01 from school age (P-N-01 had ceiling and P-N-02, P-OLE-01 had floor effects) led to slight improvement in Cronbach's alpha (with .79, without .80)

CATEGORY 4 HIGHLIGHTS

- **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 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

Overview

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

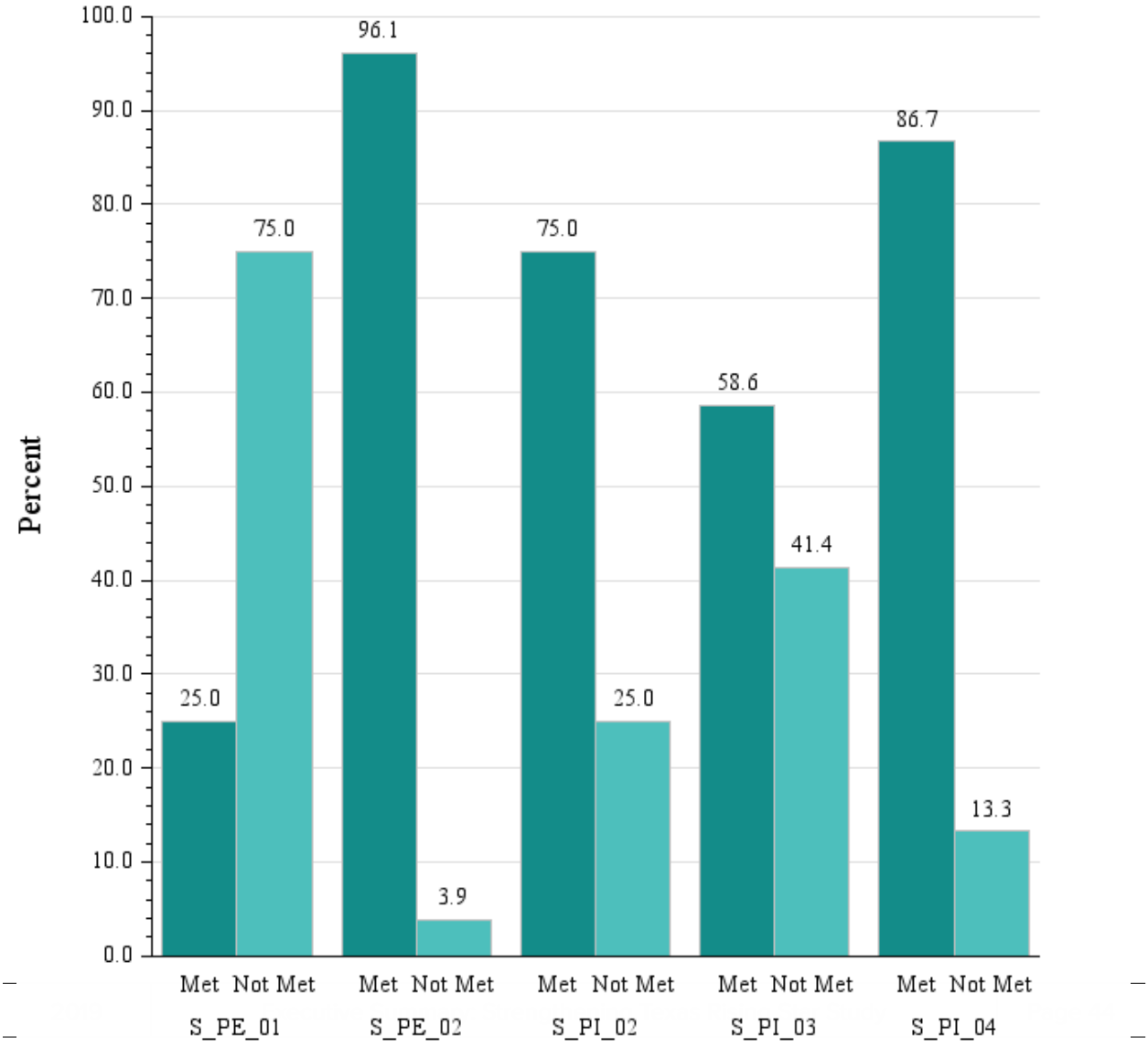
Category 5 Number of Items by Age Group

Subcategory	Number of Points-Based Items	Number of Required Items
Parent Education	2	2
Parent Involvement	3	3

Met/Not Met Items

As shown in the chart below, there is some variation among provider scores for met or not-met items. For S-PE-02, which asks Directors if they have school-parent communication systems in place, 96% of providers scored met. S-PI-04, which considers making information about community resources available to parents, was score met for 87% of the sample.

Structural Rating Percentages



Category 5 Cronbach's alpha

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.

CATEGORY 5 HIGHLIGHTS

- **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 Assessment 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.

Category 2 Factor Analysis Results

A confirmatory factor analysis (CFA) was conducted on the final items of category 2 for each of the age groups. Analytical items included LFS, PBIG, WRS, SCR (not present in Infants group), and IFAL items. Based on the model fit indices, results indicated a one-factor structure fitted data well in four age groups, meaning final items of category 2 were measuring one general construct. Moreover, in most cases, items had factor loadings larger than 0.40 across all age groups. That is, most final items of category 2 were valid to measure what they were supposed to measure. Few items had moderate factor loadings (between 0.30 to 0.40) within one or two age groups (e.g., LFS_08 had a factor loading of 0.29 in Infant group and had factor loadings larger than 0.40 in other age groups), meaning these items might be more valid in some age groups but not all groups.

Category 3 Factor Analysis Results

Original category 3 was composed of IFAL, PSNRD, and LPCand items. As presented in the recommendations table, IFAL items were moved to category 2 and PSNRD had too few items and floor effects. P_LPC items within the Pre-school group had good properties as shown in the recommendations table and therefore, were recommended to be retained for further analysis. CFA was conducted on P_LPC items for the Pre-school group. The results indicated a one-factor structure fitted data well, meaning final P_LPC items of category 3 were measuring one construct in the Pre-school group.

Furthermore, P_LPC items had factor loadings ranging from 0.56 to 0.97. More specifically, all items had factor loadings above 0.70, with the exception that LPC15C had a factor loading of

0.56. In other words, in general P_LPC items in Pro-school group were highly valid in measuring the same construct.

Category 4 Factor Analysis Results

Category 4 comprised ILE, N (not present in Infant group), and OLE items. Poor performing items were removed from the infant group and there was only one item (N_05) kept in the school-age group. Only final items were analyzed in CFA. A preliminary CFA suggested one factor model did not fit the data well. A further exploratory factor analysis (EFA) suggested ILE, N, and OLE were three distinctive factors. In other words, these three factors explicitly presented different dimensions under category 4.

To test the multi-dimensional feature of category 4, a two-factor model (ILE and OLE) was fitted to the Infant data; while a three-factor model (ILE, N, and OLE) was fitted to the remaining age groups. CFA results showed multiple-factor models fit data well across all age groups and items reasonably loaded on the factor that they belonged to (e.g., ILE items loaded on ILE factor). The results confirmed the multi-dimensional feature of category 4 existed and the results also supported items within a dimension (e.g., ILE) were measuring one dimension.

The factor loadings and correlations between dimensions/factors are presented as follows. Infants. Items' factor loadings ranged from 0.51 to 0.96. The correlation between ILE and OLE is .45 Toddler. Items' factor loadings ranged from 0.39 to 0.90. The correlations between factors were .53 (between ILE and N), .45 (between ILE and OLE), and .45 (between N and OLE). Pre-school. Items' factor loadings ranged from 0.43 to 0.82. The correlations between factors were .63 (between ILE and N), .46 (between ILE and OLE), and .30 (between N and OLE). School-age. Items' factor loadings ranged from 0.56 to 0.89. The correlations between factors were 0.24 (between ILE and N), 0.33 (between ILE and OLE), and 0.16 (between N and OLE). Based on these results, in general items within each of dimensions were highly valid in measuring the dimension that they belonged to. On the other hand, the results also showed the magnitudes of correlations between dimensions were small to moderate, meaning dimensions were distinctive. This feature is different from what we found in other categories. A theoretical framework or empirical evidence will be needed to determine whether these dimensions were under a more general factor.

Category 5 Factor Analysis Results

Items of category 5 included PE and PI items. CFA results showed a one-factor structure fitted data well, suggesting final items of category 5 were measuring one construct. Items had factor loadings ranging from 0.59 to 0.71. That is, final items were highly valid in measuring the same construct.

Overall Internal Consistency for Points-Based Items across Categories 2, 3, and 4

We examined Cronbach's alpha for points-based items using the current TRS measure structure and the recommended measure structure and found small improvements across several categories and age groups. Results by category are shown in the following table.

Overall Internal Consistency for Points-Based Items across Categories 2, 3, and 4

Category	Age Group	Current	Recommended
Category 2	Infants	0.90	0.93
	Toddlers	0.91	0.94
	Preschool	0.91	0.94
	School-age	0.90	0.92
Category 3	Infants	0.66	None tested
	Toddlers	0.60	None tested
	Preschool	0.85	0.91
	School-age	0.51	None tested
Category 4	Infants	0.60	0.79
	Toddlers	0.80	0.80
	Preschool	0.79	0.79
	School-age	0.79	0.80
Overall	Infants	0.87	0.92
	Toddlers	0.91	0.93
	Preschool	0.91	0.93
	School-age	0.90	0.91

Overall Inter-Rater Agreement with Current and Recommended Measure Structure

Generalizability coefficient was estimated for the 10 raters released for independent classroom assessment. G-coefficient was estimated overall for all points-based, classroom-level items in category 2, 3, and 4 for the current and alternate scoring methods, with rater-level reliability ranging from .67 to .89 (current scoring) and .71 to .89 (alternate scoring). Generalizability coefficients were slightly higher for the alternate items: of the 10 raters released, one rater reached reliability in the excellent range (.9), five raters achieved reliability in the good range, and three reached reliability in the acceptable range. One rater failed to maintain reliability and was reassigned.

We also examined generalizability coefficients under the new measure structure. G-coefficient was estimated overall for all points-based, classroom-level items in category 2, 3, and 4 for the current and alternate scoring methods, with rater-level reliability ranging from .73 to .90. Generalizability coefficients were slightly higher with the new structure, with 8 of the 9 raters in the "good" to "excellent" range (one rater remained in the "acceptable" range). 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 by Category

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 likely that providers could meet these requirements using standardized templates and sample documents.

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 following table reflects distributions of star ratings by category based on points-based items only.

Distribution of Star Ratings Overall and 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

The sample included 69 providers who were certified prior to or during the data collection phase of the TRS study. The following table shows that TRS providers on average had higher scores across all categories. Controlling for SES, the differences between TRS and non-TRS providers are statistically significant for categories 2, 4, and 5 ($p<.01$).

TRS Participation	Category	N	Mean	Std Dev	Min	Max
No	Average_ Category_1	59	1.23	0.34	0.20	2.00
	Average_ Category_2	59	1.30	0.38	0.41	2.17
	Average_ Category_3	59	0.36	0.25	0.02	1.54
	Average_ Category_4	59	1.18	0.44	0.46	2.20
	Average_ Category_5	59	1.25	0.70	0.20	3.00
Yes	Average_ Category_1	69	1.30	0.41	0.00	2.40
	Average_ Category_2	69	1.47	0.33	0.61	2.20
	Average_ Category_3	69	0.53	0.37	0.09	1.68
	Average_ Category_4	69	1.49	0.34	0.74	2.23
	Average_ Category_5	69	1.70	0.74	0.00	3.00

Star Rating Distributions under Recommended Structure

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.

Alternate Calculation of Star-Rating for Study Sample

The study examined whether an alternate star-level scoring based on average scores (rather than median scores) led to slightly different star rating by category but same overall rating. The study did not find significant changes to scores, therefore there is no evidence to support altering the current scoring method.

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 measures 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?
- Is there evidence that star ratings and classroom quality vary by socioeconomic status?
- Is accreditation related to TRS scores?
- Do directors with higher levels of education, training, and experience have higher scores on TRS facility items?
- Do caregivers with higher levels of education, training, and experience have higher scores in caregiving behaviors?
- Do lower child-caregiver ratios relate to higher TRS scores?
- Do TRS scores for caregiving behavior (e.g., category 2) relate to another established measure of caregiving quality (convergent validity)?
- Is the TRS assessment sensitive to changes in caregiver-child interaction quality associated with quality improvement efforts?

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 rating 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. There were no changes in overall star ratings or in ratings for categories 1 and 3. However, several facilities experienced a change in rating within categories 2 (3 facilities), 4 (6 facilities), and 5 (2 facilities). Only one facility experienced a change in rating between Visit 2 and 3 for category 4. 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). See following table for changes in score over time. 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 (observation 1 to 2, $p < .01$; observations 2 to 3, $p < .05$). Differences in scores for categories 3 and 4 were not statistically significant over time.

Stability of Ratings at the Classroom Level

Stability of Ratings across Time												
Category	Visit 1				Visit 2				Visit 3			
	N	Mean	STD	Range	N	Mean	STD	Range	N	Mean	STD	Range
2	269	37.3	12.9	2 - 70	269	34.2	12.8	0 - 63	105	31.7	12.8	5 - 62
3	269	8.7	6.5	0 - 36	269	8.2	6.7	0 - 40	105	8.5	6.4	0 - 32
4	269	19.6	6.0	3 - 34	269	19.6	6.3	4 - 34	105	19.6	5.8	4 - 31

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. The following tables show differences in scores over time.

Category Scores for Classes with all Teachers the Same across Three Visits

Visit #1					
Variable	N	Mean	STD	MIN	MAX
Category2_CTR	40	38.2	12.7	10	59
Category3_CTR	40	7.8	3.8	1	16
Category4_CTR	40	19.4	6.6	5	29

Visit #2					
Variable	N	Mean	STD	MIN	MAX
Category2_CTR	40	35.4	13.2	8	63
Category3_CTR	40	8.7	6.1	0	32
Category4_CTR	40	20.7	6.6	7	34

Visit #3					
Variable	N	Mean	STD	MIN	MAX
Category2_CTR	40	32.0	13.2	11.0	62
Category3_CTR	40	8.5	6.6	0.0	32
Category4_CTR	40	20.7	6.5	4.0	31

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 ($P < .01$ between assessment 1 and 2, and $P < .05$ between assessment 2 and 3). 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 (Curby, Grimm, & Pianta, 2010; Hill, Charalambous, & Kraft, 2012; Malmberg, Hagger, Burn, Mutton, & Colls, 2010; Mantzicopoulos, French, Patrick, Watson, & Ahn, 2018; Plank & Condliffe, 2013), 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.

Category 1

Most items in this category were scored as not met, and in general there were no differences by SES. There was some variation in the finding for TRS orientation, with middle SES providers being more likely to score met (11.3%). Mid and higher SES programs are also more likely to meet the indicator for substitute and volunteer orientation, 20% and 29% respectively. This could suggest that lower SES programs are less likely to work with volunteers than providers in higher SES communities.

Distribution of Category 1 Point-based items by SES

Item	Scoring	SES	N	Mean	Std Dev	Min	Max
P_CQT_01	#of caregivers meeting one of qualifications*100/ total # of caregivers=% of staff.(0= 0%-29%, 1=30%-50%, 2=51%-74%, 3=75% or more of staff)	Low	40	0.18	0.55	0	3
		Medium	53	0.26	0.68	0	3
		High	35	0.29	0.62	0	2
P_CQT_03	#of training topics aligned with core competencies*100/ divided by total number of training topics=% of training aligned (0= 0%-49%, 1= 50%-64%, 2= 5%-79%, 3= 80% or more of the training aligned with core competencies)	Low	40	2.50	1.11	0	3
		Medium	53	2.72	0.89	0	3
		High	35	2.89	0.40	1	3

Item	Scoring	SES	N	Mean	Std Dev	Min	Max
P_ DEQT_01	Formal Education scoring	Low	40	0.58	1.03	0	3
		Medium	53	0.81	1.06	0	3
		High	35	0.74	1.01	0	3
P_ DEQT_04	0=None, 1=2 years, 2=3 years, 3=4 or more years of experience in early childhood	Low	40	2.33	1.19	0	3
		Medium	53	2.49	1.03	0	3
		High	35	2.51	1.01	0	3
P_ DEQT_06	0=None, 1=Director has 36hrs, 6hrs in program admin, management & supervision, 2=Director has 36hrs, 6hrs in program admin & 3hrs in Infant/ Toddler or Pre-K guideline,3=Director has 36hrs,6hrs in program admin & 3hrs in Infant/Toddler & 3hrs in Pre-K guideline	Low	39	0.15	0.54	0	3
		Medium	50	0.20	0.57	0	3
		High	34	0.18	0.39	0	1

Category 2

Within our sample, the average ratio (based on children and caregivers present during the assessment) does not vary significantly by SES. The table below shows the average ratios by age group and SES. Note: Within the infant age group, some classrooms may be staffed based on a 2:10 ratio, which is not accounted for.

Average Group Size Ratio by SES

Age Group	Low SES	Medium SES	High SES
Infant	1:3	1:3	1:4
Toddler	1:7	1:6	1:6
Preschool	1:9	1:9	1:9
School-age	1:13	1:14	1:14

We also examined for differences in category 2 point-based scores. For the average of median scores (i.e., current TRS scoring procedure), there is a slight trend toward higher scores within higher SES providers, (mean scores are: high SES=1.47, middle SES=1.39, low SES=1.32. It is

important to note 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 (1.8 and 2.4, respectively).

Category 3

For category 3 point-based scores, the average of median scores (i.e., current TRS scoring procedure) shows a slight trend toward higher scores within higher SES providers, (mean scores are: high SES=.51, medium SES=.44, low SES=.43. It is important to note 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 (1.8 and 2.4, respectively).

Category 4

Category 4 shows variation across met and not met items, and there are some differences by SES. In most cases, high SES programs are more likely to meet category 4 structural quality indicators (items 1-3 below). However, for one item, S-N-03, providers in low income areas are more likely to meet the indicator. These providers may be more likely to participate in the Child and Adult Care Food Program (one way to meet item criteria).

- S-ILE-01: Measure related to the indoor environment for all ages (e.g., facilitates a distinct division of active and quiet spaces; five total criteria in item)
- S-ILE-02: Measure related to the indoor environment for infants (e.g., sufficient quantity of sleeping, diapering, and feeding equipment; four total criteria in item)
- S-N-01: Measure related to written program nutrition policies for all ages (e.g., healthy snacks are available; five total indicators)
- S-N-03: Measure related to menu planning for all ages

For category 4 point-based scores, the average of median scores (i.e., current TRS scoring procedure) shows a slight trend toward higher scores within higher SES providers, (mean scores are: high SES=1.41, medium SES=1.41, low SES=1.22. It is important to note 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 (1.8 and 2.4, respectively).

Category 5

Within category 5 there are only two items that vary slightly by SES, with programs serving higher SES families being more likely to receive child growth and development information from the provider (P-PE-02) and to receive information about their child's experience, which may include written documentation (S-PI-03).

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 measures. 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. Of the 18 providers assessed most scored at a 2-Star level, with the following exceptions:

- 6 scored at a 3-star rating in category 2
- 7 scored at a 3-star in category 4
- 5 scored at a 3-star and 5 scored at 4-star in category 5

The tables below show score by national accreditation type: National Accreditation Commission for Early Care and Education Programs (NAC), National Association for the Education of Young Children (NAEYC), and AdvancED Quality Early Learning Standards (QELS).

Category 1

Accredited By	2 Star	3 Star	4 Star	Total
NAC	9 (100%)	0 (0%)	0 (0%)	9
NAEYC	3 (75%)	1 (25%)	0 (0%)	4
QELS	5 (100%)	0 (0%)	0 (0%)	5
Total	17 (94%)	1 (6%)	0 (0%)	18

Category 2

Accredited By	2 Star	3 Star	4 Star	Total
NAC	5 (55.6%)	4 (44.4%)	0 (0%)	9
NAEYC	2 (50%)	2 (50%)	0 (0%)	4
QELS	5 (100%)	0 (0%)	0 (0%)	5
Total	12 (66.7%)	6 (33.3%)	0 (0%)	18

Category 3

Accredited By	2 Star	Total
NAC	9 (100%)	9
NAEYC	4 (100%)	4
QELS	5 (100%)	5
Total	18 (100%)	18

Category 4

Accredited By	2 Star	3 Star	4 Star	Total
NAC	6 (66.7%)	3 (33.3%)	0 (0%)	9
NAEYC	1 (25%)	3 (75%)	0 (0%)	4
QELS	4 (80%)	1 (20%)	0 (0%)	5
Total	11 (61.1%)	7 (38.9%)	0 (0%)	18

Category 5

Accredited By	2 Star	3 Star	4 Star	Total
NAC	5 (55.6%)	2 (22.2%)	2 (22.2%)	9
NAEYC	1 (25%)	0 (0%)	3 (75%)	4
QELS	2 (40%)	3 (60%)	0 (0%)	5
Total	8 (44.4%)	5 (27.8%)	5 (27.8%)	18

Based on this small sample of providers we did not find evidence to support automatic 4-star ratings for nationally accredited programs.

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

We examined correlations between all category 1 Director-focused items and TRS classroom measures and found no consistent patterns. We found a significant small correlation for Director Qualifications P-DEQT-01 and category 2 average score ($r = .22$, $p < .05$)

Given that TRS qualifications items are scored based on combinations of many indicators, we also looked at the extent to which individual indicators relate to classroom and facility points at the category level. We found multiple small to moderate significant correlations with facility-focused categories, shown below. This suggests information is lost with the current item structure, which may limit predictive validity.

Correlations of Director Qualification with Categories 1, 4, and 5 Scores

Category	Years of Experience	Business Mgmt Training Hours	Child Care Related Training Hours	Program Admin Training Hours	Highest Level of Education Achieved
Category 1	.29*	.29*	.11	.20*	.34*
Category 2	.19*	.08	.14	.04	.12
Category 3	.06	.03	.11	-.02	.12
Category 4	.10	.22*	.04	.15	.09
Category 5	0	.20*	.22*	.21*	.02

(p*<.01)

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

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

- Providers with more qualified staff (P-CQT-01) have higher scores on category 2 (r= .26, p< .05) and 4 measures (r= .45, p< .01), and category 4 star rating (r= .62, p< .01)
- Caregiver staff training topic alignment (P-CQT-03) is moderately related to category 3 scores (r= .42, p< .01).

We examined for evidence that specific indicators (i.e., elements within measures) of caregiver training and experience related to other quality measures within the TRS.

Because the classroom-level measures consider all caregivers (i.e., score may not reflect the behavior of an individual), and in order to look at the strength of association between education and caregiving practice, we looked at a subsample of classrooms with only one caregiver n=420). Within this group we found several small but significant positive correlations between education and classroom quality scores, shown below. The percentage of staff who completed pre-K and infant and toddler guidelines trainings was too low (<2%) to report correlations for these indicators.

Correlations of Caregiver Qualifications with Categories 2 and 3 Total Scores

Category Total Score	Years of Experience	Clock Hours Last 5 years	Valid CDA	Highest Level of Education Achieved	Hours of Child Care Related Training
Category 2 Total Score	.13*		.26**	.03	.17**
Category 3 Total Score	.12*	.25**	.22**	.11	.23**

(*p<.05) (**p< .01)

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

Low child-caregiver 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 (r = -.19 and .16 respectively; p < .01). When we looked at actual ratio by age group (shown in the table below) we found that correlations were small across age groups.

Correlation between Ratio and Categories 2, 3, and 4

Category	Age Group	GroupRatio_ Present
Category 2	Infants	-0.20*
	Toddlers	-0.19*
	Preschool	-0.17*
	School-age	-0.13
Category 3	Infants	0.003
	Toddlers	-0.09
	Preschool	0.13*
	School-age	-0.01

Category	Age Group	GroupRatio_ Present
Category 4	Infants	0.06
	Toddlers	0.04
	Preschool	0.004
	School-age	0.18*

$p<.05$

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 other established measures of caregiver interaction quality, the Arnett Caregiver Interaction Scale. Raters scored the Arnett and TRS classroom measures during the same observation period (n=495). The table below shows multiple high significant correlations with category 2 scores. Category 3 also includes behavioral items that should relate to Arnett constructs, and at the category level, we do see significant small to moderate correlations. Given that not all items within category 3 relate to caregiving behavior, we looked at patterns of correlation at the sub measure level, and found higher correlations with Instructional Formats and Approaches to Learning than with non-behavioral items. These data provide initial evidence that the behavioral measures within the TRS assessment relate well to other measures in routine use.

Arnett Correlations with Category 2 and Category 3 Total Scores

Arnett Subscale	Category 2	Category 3
Detachment	.72	.40
Permissiveness	.38	.27
Positive Relationships	.80	.39
Harshness	.63	.20
Total Score	.81	.37

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

Using data collected from a random assignment pilot study funded by private foundations, we also examined for evidence of TRS category 2 (Caregiver-Child Interactions) external validity. The pilot study was the initial evaluation of an educational intervention developed to

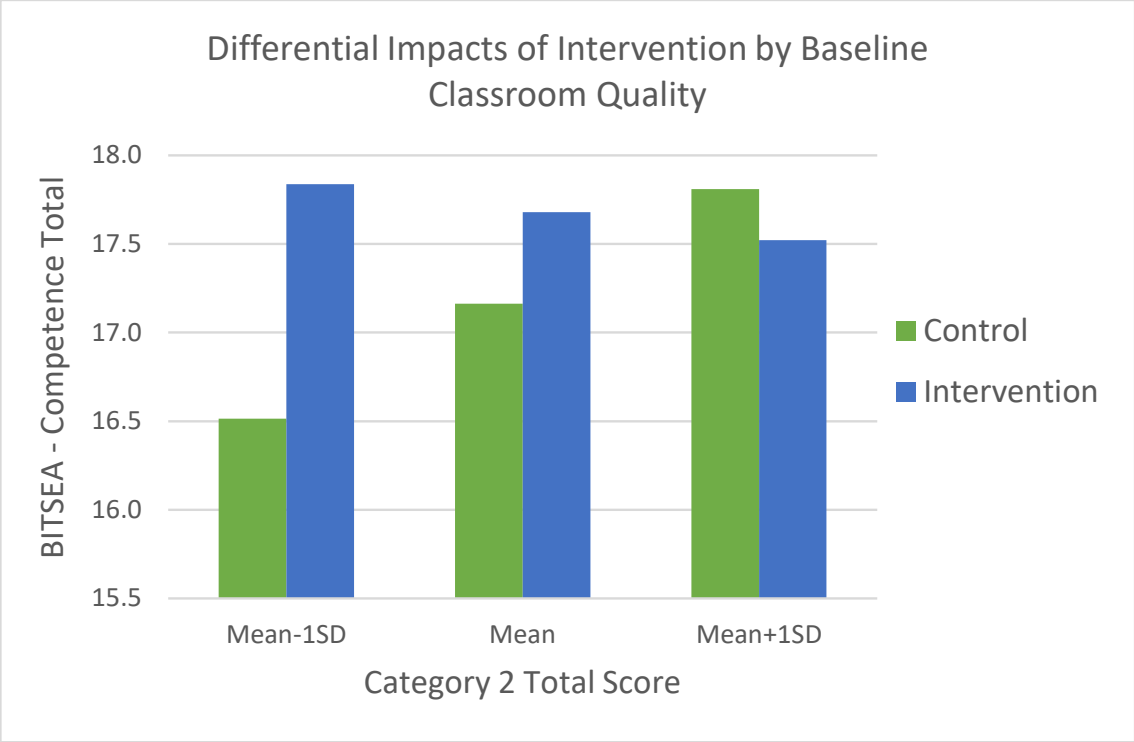
support childcare providers, CIRCLE Infant & Toddler Teacher Training: Play with Me. The purpose was to examine: 1) the feasibility of the program and impact on the quality of toddler teachers' instruction and child outcomes; 2) use of individualized coaching to support teachers' professional learning within the program; and 3) use of child progress monitoring measures (milestones checklists) to identify children who need additional support. This cluster randomized trial (CRT) occured in 40 toddler classrooms and enrolled up to 6 children per classroom (ages 24-36 months).

The Toddler Pilot sample included 38 teachers in Dallas and Houston (18=intervention, 20=control). Participating pilot teachers had an average of 6 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. Training was offered to control teachers at the conclusion of the study. Sites for the pilot study were jointly recruited into the TRS study to allow for alignment between TRS data collection and study outcomes data in a subsample of classrooms. 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 ($p<.01$, $d=.69$). 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 ($p=.003$, $d=.42$) and warm and responsive style ($p=.04$, $d=.65$). 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 ($b=-.046$, $p=.025$). As illustrated in the graph, 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.

The study employed recruitment procedures designed to maximize sample variation (e.g, urbanicity, SES), increasing our chances of observing a full range of quality across TRS items (i.e., low and high scores). While variation at the classroom level was sufficient to allow us to explore reliability research aims (e.g., some caregivers exhibited high quality behaviors), star-level calculation procedures resulted in very few providers reaching 3- or 4-star quality at the category level. This low variation limited our ability to examine relations between overall star ratings and other outcomes (external validity).

Finally, our initial exploration of validity was limited given our primary focus on reliability and 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 than what was possible in the scope of this study.

Section 4

Recommendations

We recognize that there may be multiple goals for quality rating and improvement systems (QRIS), for example, 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.

The following table summarizes key rationale(s) applicable to each recommendation, and include:

- **Reliability** covers a range of concerns that were discussed in detail in the results section, and may include concerns about item-level functioning (e.g., ceiling effects), frequent exclusion from scoring, internal consistency, or inter-rater agreement
- **Validity** concerns relate to the extent to which the assessment functions as expected and is supported by evidence (e.g., accreditation rules).
- **Training** concerns refer to the influence of significant barriers to achieving reliability or for supporting proper implementation that should be attended to and inform the recommendation (e.g., item required extensive training and resulted in low agreement).
- **Implementation** concerns refer to a range of factors we believe limit the usefulness of items or measures within the assessment, and should inform changes to current practice (e.g., lengthy scoring time, inconsistent access to information).

In some cases these rationales are interrelated and/or converge. For example, some items do not contribute to a measure's reliability, are difficult to train, are scored based on information from providers that is inconsistently available, and may lack external evidence of their importance toward reaching one or more of the QRIS goals described above.

Recommendation	Reliability	Validity	Training	Implementation
Recommendation 1: Removing or adjusting low-performing items to improve instrument functioning	•		•	•
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	•	•		
Recommendation 3: Revising procedures for automatic certification of nationally accredited providers		•		
Recommendation 4: Employ a rigorous training and reliability monitoring process to ensure accurate star rating across the state	•	•	•	•
Recommendation 5: Standardizing application and scoring routines to improve program efficiency and accuracy of star assignment	•		•	•
Recommendation 6: Establishing a quality improvement framework that uses a developmental approach to ensure providers receive technical assistance and professional development in alignment with their current star ratings	•			•
Recommendation 7: Continuing exploration of external validity		•		

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. Please refer to Table 1 in the Recommendations for Item Revision or Removal tables in Appendix 4 for a list of these items.

Table 2 in Appendix 4 includes the remaining 29% of items, for which 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 results section. 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. There has been an increasing focus on understanding the process of implementing high-quality curricula rather than focusing solely on the curricula itself (Daily, Hegseth, & Michael, 2015). That is, understanding how curricula are implemented is an essential, but often missing, component of many QRIS curricula quality indicators (Chazan-Cohen et al., 2017). Generally, high quality implementation of curricula includes using developmentally appropriate materials, engaging in iterative cycles of assessment, documentation, and planning based on young children's interests, and delivering curricula within contexts of nurturing and responsive caregiver-child relationships (Chazan-Cohen et al., 2017). These 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 will also yield data 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: Adjust the relative weight of structural vs. process measures 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 (e.g., director qualifications) (e.g., Hamre & Pianta, 2005; Howes et al. 2008; Mashburn et al., 2008). 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 base certification ratings on full site assessment scores. Having accurate information about program quality that is specific to the TRS standards 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" to "excellent" inter-rater reliability after a rigorous training process. 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 state with incredible diversity, 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.

The Strengthening Texas Rising Star Implementation Study included the design and development of the TRS Assessment Training and Certification Program. The program includes online learning modules, practice assignments, and a tiered support approach for staff who do not meet reliability criteria, including small group PLCs and individualized feedback. CLI has completed a considerable portion of certification program for all categories (e.g., collection of document artifacts used to train and as a basis for practice assignments). A description of the program can be found in Appendix 9. Finalization of the certification system is pending commission decision on changes related to the Texas Rising four-year program review.

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 note taking and documentation procedures to help bring clarity to the ratings process, strengthen the accuracy of ratings, and provide evidence for specific scores in communications with providers. This is particularly important for category 2, 3, and 4. The specific notetaking forms and worksheets used in the study can be found in Appendix 6, however, we recommend these be combined in a redesign of the assessment record forms.

There is also an opportunity to increase rating efficiency and accuracy of category 1, which 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. The Texas Early Childhood Professional Development System (TECPDS) Workforce Registry includes individual staff reports that detail their education, qualifications, and training that can be used to facilitate scoring of director and caregiver qualifications in category 1 (for a list of category 1 indicators that could be captured by TECPDS, see Appendix 7). When TECPDS was used to facilitate scoring in study, time estimates dropped to 10-15 minutes per caregiver/director. This procedure was also successfully piloted with the Tarrant County region with TRS and non-TRS providers across the LWDB region in a separate project. In addition to Tarrant County, 11 other LWDBs are currently in the TECPDS onboarding process. We also recommend Integrating TECPDS with the TRS 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.

The depth and breadth of TRS standards and the individual needs of providers make individualizing technical assistance a challenging endeavor. Implementation science provides a helpful conceptual framework to organize the types of technical assistance TRS might offer providers depending on the maturity of their TRS participation and their current levels of quality. Metz and Bartley (2012) describe four integral stages of implementation: (1) **exploration**, which describes collecting contextual information (e.g., resources, staff) to determine feasibility; (2) **installation**, which describes the process of setting up the practical and logistic aspects of implementation (e.g., training providers, purchasing needs); (3) **initial implementation**, which describes the process of using rapid problem solving and data driven approaches to assess and improve implementation efforts; and (4) **full implementation**, which describes the process of enacting stable procedures that facilitate implementation of the new program. Below, the authors offer potential technical assistance activities aligned with these stages that promote eventual 4-star certification and sustained high quality. Each stage uses virtual professional learning communities (PLCs) that provide predictable support windows and achieve information dissemination, offer opportunities for provider Q&A, and facilitate peer-to-peer support. More intensive technical assistance support would be provided in later stages, such as the development of continuous quality improvement (CQI) plans (stage 3) and individualized coaching (stage 4). A summary CQI graphic can be found in Appendix 5.

Stage 1: Awareness / Interest in TRS (Exploration)

To promote interest and awareness in the Texas Rising Star certification process, TRS could expand its public-facing resources to include overview videos/documents, video exemplars of key indicators (e.g., caregiver-child interactions), sample templates that demonstrate how to align plans to TRS standards (e.g., caregiver training plans), and consumer education materials to support family enrollment. TRS could also host a recurring "Introduction to TRS" PLC that includes discussion of the guidelines, certification process, and opportunities for technical assistance. These materials and PLC opportunities would be actively disseminated by TRS staff to providers in their local communities.

Stage 2: Self-Assessment and TRS Application (Installation)

This stage includes orientation activities such as completing the TRS orientation, onboarding to the TECPDS (which the authors propose can be used to streamline scoring of specific structural indicators), and how to access quality improvement resources (such as CLI Engage that offers free resources). A recurring PLC, "How to Complete the TRS Self-Assessment," would provide support to providers on what to look for as they complete their self-assessment. Once the provider has completed the self-assessment, an automated report could be generated that shows national, state, and local resources aligned to specific sub-categories. Local resources could include community-based trainings that tightly align with the TRS categories (supported by the TECPDS Trainer Registry). Based on their self-ratings, these resources could be used by the provider to support initial quality improvement efforts, such as making improvements to indoor and outdoor learning environments. Given that very few providers (see page 28), have completed state guidelines training (infant and toddler, prekindergarten), the authors recommend these trainings be included in an initial set of required CQI activities.

Stage 3: TRS Assessment (Initial Implementation)

The TRS assessment occurs in the initial implementation phase, after which both providers and TRS Mentors are jointly defining needs and creating improvement plans. The authors recommend that continuous improvement plans be established regardless of whether the assessment results in certification (i.e., plan still generated if the provider is not certified but remains committed to certification). A recurring PLC for providers at this stage could focus on how CQI is used as an approach to make incremental improvements in areas of the TRS guidelines. Communicating clearly about CQI is likely critical for motivating providers who were not certified to continue to engage with the program.

Stage 4: Continuous Quality Improvement Activities (Full Implementation)

During full implementation, providers are fully immersed in the TRS program through opportunities to participate in ongoing topic-based PLCs (e.g., infant and toddler language development) and individualized CQI efforts. The CQI framework draws upon a cycle in which providers and staff (1) **assess** their practice and children's learning, (2) **set goals** based on needs, (3) **enact a plan** for practicing concepts and/or strategies, (4) receive expert **feedback**, and (5) **reflect** on their progress as they begin another improvement cycle. These cycles are designed to achieve incremental improvements in realistic timelines. The goal-setting process prioritizes TRS indicators that are not at a 4-star level (facilitated by automated reporting features). As mentioned, action plans for achieving goals would include leveraging a variety of national, state, and local resources. To facilitate these cycles, the authors recommend that CQI efforts include specialized PLCs and individualized coaching of directors and classroom-level staff most in need of intensive support. In both home-based and center-based child care, the use of coaching supports has been found to positively influence the quality of caregiving (Shivers, Farago, & Goubeaux, 2015). To expand TRS participation and mentoring support to new communities in need, remote (virtual, video-based) coaching models should be strongly considered. Formal

monitoring of CQI plans can also inform performance prior to annual monitoring visits as well as readiness for star-level assessment or reassessment (e.g., strong progress in completing CQI action plans may prompt an assessment for non-certified providers). This also helps incentivize CQI activities for providers working toward certification or higher reimbursements. Although CQI is designed to help providers progress through star levels to achieve the highest quality rating, CQI activities should not cease when a 4-star rating is awarded. As the goal is to reach and sustain high quality, 4-star-rated providers would continue CQI activities as needed, for example when new staff need onboarding and coaching.

Role of TRS Mentors

The above recommendations require TRS Mentors to implement the quality improvement framework in each stage, from exploration (sending out public-facing onboarding materials, first-line technical assistance) to full implementation (mentor-supported continuous improvement planning and targeting of coaching to those plans). This will require training for mentors to effectively implement these strategies.

CQI Planning

While the above recommendations are based on significant research evidence and implementation expertise, implementing a high quality technical assistance system such as the one described above will require an intensive planning period that brings together TRS stakeholders to adapt available CQI models to align with the TRS standards and dive deeper into community needs and resources that can be leveraged.

Recommendation 7: Continued exploration of external validity in alignment with QRIS goals, including long-term rating stability and evidence of impacts on outcomes of interest.

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. We found some initial evidence of validity (e.g., strong correlations between TRS caregiver-child interactions and validated measures of caregiving quality) by examining data in our study sample. Once statewide field reliability is established, 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).

References

About QRIS. (n.d.). Retrieved from https://qrisguide.acf.hhs.gov/about-qris.

Adams, G., Zaslow, M., & Tout, K. (2007). Early care and education for children in low income families: Patterns of use, quality, and potential policy implications. *The Urban Institute in Child Trends*.

Booth, C. L., & Kelly, J. F. (2002). Child care effects on the development of toddlers with special needs. *Early Childhood Research Quarterly*, 17(2), 171—196

Burchinal, M. R., & Cryer, D. (2003). Diversity, child care quality, and developmental outcomes. *Early Childhood Research Quarterly*, 18(4), 401—426.

Briggs-Gowan, M. J., Carter, A. S., Irwin, J. R., Wachtel, K., & Cicchetti, D. V. (2004). The Brief Infant-Toddler Social and Emotional Assessment: screening for social-emotional problems and delays in competence. *Journal of Pediatric Psychology*, 29(2), 143-155.

Chazan-Cohen, R., Zaslow, M., Raikes, H. H., Elicker, J., Paulsell, D., Dean, A., & Kriener-Althen, K. (2017). Working toward a Definition of Infant/Toddler Curricula: Intentionally Furthering the Development of Individual Children within Responsive Relationships. OPRE Report 2017-15. Office of Planning, Research and Evaluation.

Colwell, N., Gordon, R. A., Fujimoto, K., Kaestner, R., & Korenman, S. (2013). New evidence on the validity of the Arnett Caregiver Interaction Scale: Results from the early childhood longitudinal study-birth cohort. *Early Childhood Research Quarterly*, 28(2), 218—233.

Curby, T. W., Grimm, K. J., & Pianta, R. C. (2010). Stability and change in early childhood classroom interactions during the first two hours of a day. *Early Childhood Research Quarterly*, 25(3), 37—84.

Daily, S., Hegseth, D., & Michael, L. R. (July 16, 2015). Measuring curriculum & assessment implementation in QRIS. Paper presented at the QRIS National Meeting, National Harbor, MD. Retrieved from https://qrisnetwork.org/sites/all/files/conference-session/resources/Session231PPTasPDF.pdf

Duncan, G. J., & Murnane, R. J. (Eds.). (2011). *Whither opportunity?: Rising inequality, schools, and children's life chances*. Russell Sage Foundation.

Duncan, G. J., Morris, P. A., & Rodrigues, C. (2011). Does money really matter? Estimating impacts of family income on young children's achievement with data from random-assignment experiments. *Developmental Psychology*, 47(5), 1263—1279.

Fontaine, N. S., Torre, D. L., & Grafwallner, R. (2006). Effects of quality early care on school readiness skills of children at risk. *Early Child Development and Care*, 176(1), 99—109.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, 76(5), 949-967.

Hanushek, E. A., Peterson, P. E., Talpey, L. M., & Woessmann, L. (2019). *The Unwavering SES Achievement Gap: Trends in US Student Performance* (No. w25648). National Bureau of Economic Research.

H.B 376, 83rd Texas Legislature 2013 Reg.Sess (Texas, 2013)

Heckman, J. J., & Karapakula, G. (2019). *The Perry Preschoolers at late midlife: A study in design-specific inference* (No. w25888). National Bureau of Economic Research.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56—64.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, 23(1), 27-50.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.

Magnuson, K., & Waldfogel, J. (Eds.). (2008). *Steady gains and stalled progress: Inequality and the Black-White test score gap*. Russell Sage Foundation.

Malmberg, L. E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, 102(4), 916—932.

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: a generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment*, 23(1), 24-46.

Marcoulides, G. A. (2000). Generalizability theory. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527—551). Academic Press.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., ... & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development*, 79(3), 732-749.

Metz, A., & Bartley, L. (2012). Active Implementation Frameworks for Program Success: How to Use Implementation Science to Improve Outcomes for Children. *Zero to Three* (J), 32(4), 11—18.

National Center on Early Childhood Quality Assurance (2013). *QRIS Elements*. Retrieved from: https://childcareta.acf.hhs.gov/sites/default/files/public/254_1302_qris_elements.pdf

Plank, S. B., & Condliffe, B. F. (2013). Pressures of the season: An examination of classroom quality and high-stakes accountability. *American Educational Research Journal*, 50(5), 1152—1182.

Roberts, J., Marshall, N. L., Tracy, A., Santaniello, S., Melia, M., Moore, H.,...Khlifi, K. (2011). Massachusetts Quality Rating and Improvement System (QRIS) Validation Study: Final Report. Retrieved from: https://www.mass.gov/files/2017-08/Revised%20Validation%20Study%20ReportfinalFORMATTED.pdf

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44(6), 922—932. http://dx.doi.org/10.1037/0003-066X.44.6.922

Shivers, E. M., Farago, F., & Goubeaux, P. (2015). *The Arizona Kith and Kin Evaluation, Brief #1: Improving quality in family, friend, and neighbor (FFN) child care settings.* Indigo Cultural Center, for the Association for Supportive Child Care, with support from First Things First and Valley of the Sun United Way.

Appendix

The following pages include:

Appendix 1: TRS-Selected Deficiencies in Study Sample

Appendix 2: Item-Level Descriptives for Points-Based and Met/Not Met Items

Appendix 3: Scores by Socioeconomic Status

Appendix 4: Recommendations for Item Removal or Revision tables

Appendix 5: Continuous Quality Improvement (CQI) graphic

Appendix 6: Sample Forms Used in the Study: Facility Assessment Record Form (FARF), Classroom Assessment Record Form (CARF), Note-taking Form, and Director and Caregiver Worksheets

Appendix 7: Data Sources for Capturing Category 1 Indicators

Appendix 8: Sample Individual Profile Report from the Texas Early Childhood Professional Development System Workforce Registry

Appendix 9: TRS Assessment Training and Certification Program Description