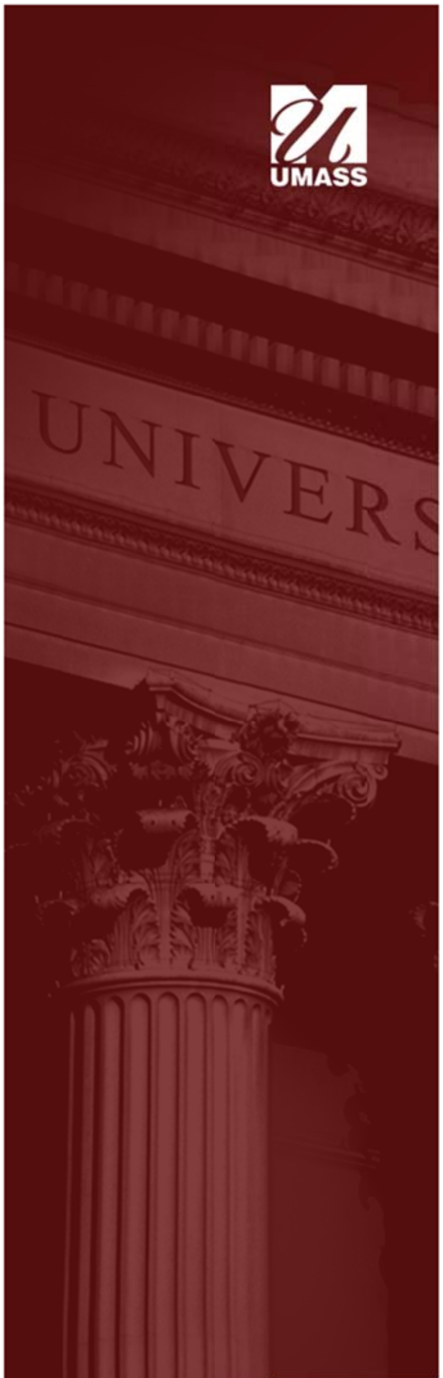
UMASS DONAHUE INSTITUTE • RESEARCH & EVALUATION GROUP



Massachusetts Quality Rating and Improvement System (QRIS) Validation Study

Study Design

A project funded by the Massachusetts Department of Early Education and Care with funding from a US Dept. of Education Race to the Top – Early Learning Challenge Grant

Revisions as of February 26, 2014

MASSACHUSETTS
Department of Early Education and Care

Validation Study Design

I. Introduction

In March 2011, the Massachusetts Department of Early Education and Care (EEC) launched its Quality Rating and Improvement System (QRIS) to assess, improve, and communicate the level of quality in early education and care and afterschool settings throughout the Commonwealth. A federal Race to the Top – Early Learning Challenge (RTTT-ELC) grant award has provided significant support to statewide implementation of the Massachusetts QRIS.

As part of the RTTT-ELC, grantees are required to conduct a study of the state's QRIS that (1) validates, using research-based measures, whether the tiers in the state's QRIS accurately reflect differential levels of program quality; and (2) assesses, using appropriate research designs and measures of progress, the extent to which changes in quality ratings are related to progress in children's learning, development, and school readiness. Not only is the current validation study designed to be responsive to key questions defined in the RTTT-ELC application, it has also been created with a keen eye toward supporting EEC's efforts for continuous program improvement. Ultimately, continuous improvement in the QRIS will maximize the quality of early education and care services to the Commonwealth's children and their families.

Although the University of Massachusetts Donahue Institute's (UMDI) at-scale study was originally slated to begin in Fall 2013, it was mutually decided that the QRIS is still in a relatively early stage of implementation and is not ready for the type of validation activity as outlined in the original research study plan approved by EEC on October 15, 2012. This document presents a draft of the revised research plan organized as follows:

II. Validation Framework

The QRIS validation framework proposed by Zellman and Fiene (2012) includes four related validation approaches and activities as follows:

1. Examine the validity of key underlying concepts,
2. Examine the measurement strategies used to assess quality,
3. Assess the outputs of the rating process, and
4. Examine how ratings are associated with children's development.

Zellman and Fiene emphasize that validation studies do not produce "yes" or "no" answers about QRIS, but do provide data that supports QRIS through identifying areas of needed refinement. As states continue implementation of QRIS, administrators and stakeholders are encouraged to undertake validation efforts that can inform their systems and to move progressively toward a more effective and efficient system. QRIS validation studies face unique challenges because it is often unknown how particular QRIS components operate and/or are related to one another, prior to the study. Additionally, these studies are vulnerable to selection biases (in programs, families, and children) and often have only modest effect sizes when exploring relations between dimensions of quality and children's development.

III. Study Design Overview

To validate the QRIS in Massachusetts, we propose a validation study that incorporates aspects of all four validation approaches outlined above. The tables below presents the research questions, major research activities and the potential policy and practice implications that may be informed by the study findings.

Validation Approach 1: Examine the Validity of Key Underlying Concepts

Research Questions	Major Research Activity	Policy/Practice Implications
1. Does the field of early education and care, nationally and in MA, support the key components of quality in the MA QRIS?	• Summarize findings from EDC report regarding QRIS standards and quality components.	• Synthesizes information from the field of early education and care and providers in MA to provide important feedback to increase program engagement.
2. What are the perceptions of the QRIS and QRIS engagement by MA programs and providers?	• Incorporate pre-validation findings from focus groups and survey of MA providers regarding QRIS.	• Provides information to inform policy regarding areas of needed outreach and community support to facilitate participation and buy-in.

Overview of Approach 1

Research Questions: Do experts in the field in MA support the key components of quality in the MA QRIS? What are the perceptions of the QRIS and QRIS engagement by programs and providers?

Prior to the proposed validation study, EEC contracted with the Education Development Center (EDC) to do a thorough review of research and literature in the field of early education and care to validate the individual QRIS components. To address this validation approach and the research questions, an overview of this work and references will be incorporated in the final report. Additionally as part of pre-validation activities, UMDI conducted focus groups with stakeholders from MA, and administered a statewide survey of approximately 600 MA center and school-based program directors and family-based providers in the field, regarding perceptions and engagement with the QRIS. If funding permits, the statewide survey will be re-administered in Fall 2015 to show changes in attitude and participation across the state over the previous 18 months. Findings from the above activities will be integrated into the final report and will supplement the EDC review, with information gathered *directly* from the field.

Policy/Practice Implications: These activities synthesize MA-specific information with the field of early education and care to provide information regarding the perceptions of QRIS and its components. It will also advance the system by identifying barriers to participation and needed support activities. It is important to note that in order for the QRIS to operate as a valid quality rating and improvement system, it must have provider buy-in, support, and participation across the state and across multiple types of programs. Therefore, these activities are particularly important for informing policy and activities related to facilitating provider outreach and engagement.

Validation Approach 2: Examine the Measurement Strategies Used to Assess Quality

Research Questions	Major Research Activity	Policy/Practice Implications
3. What are the characteristics of programs by level?	- Describe programs by level.	- Provides important information regarding the characteristics of programs at each level, highlight subgroups within each level that may need added supports to progress, and identify potential “road-blocks” at each system level.
4. Are these characteristics consistent with standards at each QRIS Level?	- Determine the reliability of programs’ Level 2, 3 and 4 ratings.	- Identifies standards and activities in which refined verification protocols would strengthen categorization within the system.
5. Do measures relate to one another as expected?	- Examine the associations among QRIS components.	- Identifies potential ways to streamline areas of overlap and redundancy.
6. Do different ways of calculating or combining scores yield more meaningful distinctions among programs?		- Identifies ways to consolidate and combine standards for enhancing the leveling of programs in a meaningful way within QRIS.

Overview of Approach 2

Research Questions: What are the characteristics of programs by level? Are these characteristics consistent with QRIS Level requirements?

These research questions are designed to describe and examine programs by level. As part of this question, the study will determine the reliability of programs’ Level 2, 3 and 4 ratings. If the rating system includes reliable and valid processes, we expect that Level 2 programs will meet the requirements for Level 2; Level 3 programs will meet the requirements for Level 3; and Level 4 programs will meet the requirements for Level 4. Rates of accuracy in classifying programs and false positives will be calculated to determine the reliability of the QRIS in classifying programs into distinct levels and the validity of QRIS assessment activities in determining the levels they achieve. Additionally, the study will describe Level 1 programs to determine the extent to which this level is diverse and includes programs that meet the standards for higher levels of the QRIS but have not applied to advance within the system.

Policy/Practice Implications: This validation activity will provide important information regarding the characteristics of programs at each level, highlight subgroups at each level that may need added supports, and identify potential "road-blocks" at each level (e.g., programs that meet all the standards for Level 3 except one, resulting in a Level 2 classification). These activities will also identify characteristics of programs at Level 1 to determine level heterogeneity and if certain program types may need additional outreach to foster advancement within the QRIS. Additionally, this activity will identity QRIS verification activities and level requirements that may need further refinement to facilitate the leveling of programs in a manner that better distinguishes quality.

Research Questions: Do measures relate to one another as expected? Do different ways of calculating or combining scores yield more meaningful distinctions among programs?

We will explore associations among various QRIS components. As part of this question, the study will examine the distribution and variation of key indicators. It will also investigate relations among and within the five quality domains of the QRIS. This question is designed to ascertain the degree to which different QRIS components are related to each other and explore differing ways to combine variables or reduce variables to assess programs' overall quality and the quality within the five QRIS quality domains.

Policy/Practice Implications: These activities will help to identify redundancies in the system in an effort to streamline the QRIS standards. In addition, the analyses will provide important information regarding ways to combine and reduce the different components of QRIS that identify meaningful distinctions in program quality. The ultimate goal of this activity is to employ data to refine standards and QRIS levels to improve effectiveness and implementation.

Validation Approach 3: Assess the Outputs of the Rating Process

Research Questions	Major Research Activity	Policy/Program Implications
7. Do programs in different QRIS levels differ significantly in observed and structural quality?	• Examine if different QRIS levels represent differences in observed and structural quality.	• Addresses RTTT requirements by determining if programs at different levels exhibit different levels of quality, or if modifications are needed to standards.
8. Do QRIS level distributions vary significantly by key program characteristics (e.g., large multi-site agencies versus single site programs)?	• Explore relations among program characteristics and QRIS level.	• Identifies program characteristics that are significantly related to QRIS level to highlight program types that may need added outreach and support.

Overview of Approach 3

Research Questions: Do programs in different QRIS levels differ significantly in observed and structural quality? Do QRIS level distributions vary significantly by key program characteristics?

We will assess the validity of the "outputs" of the QRIS, by examining whether programs in different QRIS levels differ from each other in observed and structural quality. Observational quality refers to the actual classroom experiences of children, such as children's exposure to materials and teacher-child interactions. Structural quality refers to objective aspects of the child care environment that can be regulated, such as group size, ratio, and teacher education and training. We will also explore the relations among program characteristics (e.g., large social service agencies versus single-site organizations) and QRIS levels to identify if certain program types seem to be over or under represented at each level. This will potentially identify program types that might need additional supports to progress in QRIS.

<u>Policy/Practice implications:</u> This activity addresses the requirements for RTTT funding. Additionally, it provides important information regarding whether programs at different levels exhibit significantly different levels of observed and structural quality. This serves to validate the quality level as a distinguisher of quality. These activities also explore the distribution of programs among levels to determine if certain program-types may need added supports and outreach.

Validation Approach 4: Examine How Ratings are Associated with Children's Development

Research Questions	Major Research Activity	Policy/Program Implications
9. Do children who attend higher-rated programs have greater gains than children who attend lower-level programs?	• Explore the differences in preschool children's developmental gains by the QRIS level of the program that children attend.	• Addresses RTTT requirements by examining if QRIS level are associated with outcomes in children.
10. What QRIS standards are significantly associated to increased child outcomes? 11. Does QRIS level significantly predict children's outcomes beyond key demographics of children?	• Explore differences in children's developmental gains by the specific QRIS standards of QRIS and in relation to key child demographics.	• Addresses RTTT requirements in-depth by examining which components of QRIS that are most related to outcomes. • Addresses RTTT requirements in-depth by testing if the QRIS leveling system is related to child outcomes when considering child-specific demographics

Overview of Approach 4

Research Questions: Do children who attend higher-level programs have greater gains than children who attend lower-level programs?

A central goal of QRIS is to improve child outcomes. In order to ensure that the QRIS has the capability to support improvements in programs and practice that support children's developmental progress, it is essential to demonstrate that the quality ratings are related to child growth and development in meaningful ways (Zellman & Fiene, 2012). To account for different starting points on measures of key developmental domains—such as language and literacy, early math, and social-emotional development—it is best practice to track how ratings are correlated with measures of growth or progress rather than a set benchmark or score (Zellman & Fiene, 2012). This question examines if significant differences are found in preschool children's developmental gains by the QRIS level of the program that children are attending. In other words, the study will explore whether preschool children in higher level programs exhibit significantly greater developmental gains than children attending lower QRIS level programs. It should be noted that several QRIS studies conducted by other states have not found a relationship between outcomes and QRIS. This is due in part to study limitations and the newness of the QRIS system and program ratings. The current study has been designed to enable researchers to detect smaller effect sizes and will explore associations between outcomes and QRIS level, as well as child outcomes and key components of the QRIS system.

<u>Policy/Practice Implications:</u> These activities meet the requirements for RTTT. They will assess if the QRIS levels are related to outcomes in preschool children overall and across multiple domains of development. This will help to inform whether or not QRIS is meeting its ultimate goal of improving outcomes for children.

Research Questions: What QRIS standards are significantly associated with increased child outcomes? Does QRIS level significantly predict children's outcomes beyond key demographics of children?

This question addresses Approach 4 in greater detail, and considers key demographic variables of children when examining the relations among QRIS level, QRIS components, and child outcomes.

<u>Policy/Practice:</u> These activities will further meet the requirements for RTTT. They are designed to explore which aspects of the QRIS have the strongest relationship to child outcomes. It also examines if the QRIS predicts outcomes across multiple domains above and beyond child-specific demographics. This will provide more in-depth information regarding the relationship of the different variables of the QRIS to children's outcomes and the relationship between the QRIS levels and outcomes across diverse demographic characteristics.

IV. Major Study Components

1. **Stakeholder Focus Groups:** As part of the pre-validation activities, UMDI conducted a series of regional focus groups and interviews with early education and care administrators and providers from across the state regarding the QRIS. Information gathered from these evaluation activities will be synthesized and summarized with previous work done by EEC through a contract with EDC. This will be presented as part of the final report. It will offer feedback related to validation Approach 1, from the field in general as well as from experts and providers within MA.

2. **Statewide Survey of Programs:** As part of pre-validation activities, UMDI has developed a survey for family, center, and school-based program directors regarding their experiences with QRIS, including how they get support for QRIS, plans for future participation in QRIS, additional support EEC could provide, overall opinions about QRIS and recommended changes. Data collected from this survey provides critical information regarding programs' perception of the system as a whole, of individual components of the system, and of their own participation within the system. Information from this survey will be integrated into the final report to inform questions related to validation Approach 1. If resources permit, this survey will be re-administered at the end of the study to assess changes in attitudes and perceptions that will take place over the next 18 months.

3. **Director Interview:** Created by researchers at UMDI, this structured director interview is designed to capture key program information focused on measures of quality that extend from program characteristics and policies, such that they are unlikely to differ classroom to classroom (e.g., workforce development, director qualifications, family and community engagement and leadership,

administration and management). The interview will provide important data regarding the structural, workforce, family, and leadership characteristics of quality that observations cannot capture. This will be important for defining quality and differences between levels as well as for identifying key components in which no differences are found among programs by level. Data from these interviews will be used to answer research questions related to Approaches 2 and 3.

4. **Teacher Interview:** As part of the ECERS-R and ITERS-R protocol, teachers will be interviewed to complete the scoring of specific items and the Parents and Staff subscale of the instruments. In addition to the observational interview protocol, teachers will be asked to provide information such as their job title, their highest level of education completed, the number of years of experience at their current role and their number of years of experience, trainings they have taken in the past 12 months, and questions regarding their likelihood to remain at their current position and in the field of early education and care. Data from these interviews will be used to answer research questions related to Approaches 2 and 3.

5. **Program Quality Observations:** We will collect observations in each of a random-sample of at least 120 center-based programs to provide expert assessment of program quality.

   Classroom observations will be led by Dr. Joanne Roberts of the Wellesley Centers for Women. The classroom observations will include either the ECERS-R or ITERS-R environmental rating scale, as well as the Arnett caregiver interaction scale. Dr. Roberts is certified as a reliable rater and anchor by the Environmental Rating Scales Institute (ERSI) on the ECERS-R and the ITERS-R. Dr. Roberts has also used the Arnett in multiple studies of child care quality. One classroom will be randomly selected for observation at each randomly selected preschool program and infant/toddler program. Classroom observations will take place in the winter between child assessment periods. To reduce any potential bias, observers will be blind to the level ratings of the programs. Observations will last approximately 3–4 hours and will be scheduled on a typical morning that is convenient to the program. The observers will do their best to work around the programs schedules and will not interact with the children. All indicators on the ECERS-R or ITERS-R will be rated as "yes" or "no," regardless of the final score each item. This is being done to allow of item analyses of specific indicators. Data from these observations will be used to answer research questions related to Approaches 2, 3, and 4.

Measures

ECERS-R (Harms, Clifford, Cryer, 2004): The revised ECERS-R is designed to assess process quality in preschool classrooms. It is comprised of 43 items organized into seven subscales: space and furnishings, personal care routines, language-reasoning, activities, interactions, program structure, and parents and staff. The ECERS-R is the most widely used measure of preschool classroom quality and has been used as a standard measure of quality to which other measures can be compared and validated (e.g., Burchinal, Kainz, and Cai, 2011; Bryant et al., 2003). The ECERS-R has been used in multiple studies of preschool quality including the Early Head Start Study Evaluation, the Welfare, Children and Families: A Three City Study, the Head Start FACES study, and the National Child Care Staffing Study. Multiple studies have documented a relationship between higher scores on the ECERS-R and more positive child development outcomes in areas that are considered important for later school success. The effects of higher quality early childhood experiences have now been shown to last at least through the second grade of elementary school (Peisner-Feinberg, et al., 1999).

ITERS-R (Harms, Cryer, Clifford, 2007): The Infant/Toddler Environment Rating Scale-Revised Edition (ITERS-R) is designed to assess child care quality in classrooms serving children from birth to 30 months of age, the age group that is considered most vulnerable physically, mentally, and emotionally. Therefore, the ITERS-R contains items to assess provisions in the environment for the protection of children's health and safety; appropriate stimulation through language and activities; and warm, supportive interaction. It consists of 39 items which are organized into seven subscales: space and furnishings, personal care routines, listening and talking, activities, interaction, program structure, and parents and staff. The ITERS-R has been used in multiples studies of child care quality including the Early Head Start Evaluation, and the NICHD study of Early Child Care and Youth Development.

The Global Caregiving Rating Scale (Arnett, 1989) is an observed measure of caregiver involvement and interaction style with children. It contains 26 items and has four subscales: sensitivity, harshness, detachment, and permissiveness. It has been used in numerous studies to assess the quality of the provider-caregiver relationships, including the MA Cost and Quality Studies, The ME Cost and Quality Studies, the National EHS Evaluation, and the NICHD SECCYD. Inter-rater reliability has been reported at 0.89 for the total scores (Whitebook et al., 1990) with a range from 0.92 to 0.95 for the subscales (Helburn, 1995).

6. **Child Assessment Study.** We will conduct pre- and post-observation assessments with at least 480 children in preschool classrooms and 216 toddlers in 120 programs to examine children’s developmental gains over the school year.

   We will complete pre- and post-assessments of randomly selected preschool children that are enrolled in randomly-selected preschool classrooms to determine children’s developmental gains. We will also have teachers complete pre- and post-assessments of randomly selected toddlers enrolled in the randomly selected toddler classrooms. Zellman and Fiene (2012) stress the importance of using developmental gain as a measure of outcomes as opposed to point-in-time measurements when assessing the impact of quality on outcomes, due to the limited effect size that typically is associated with this type of study. Pre-assessments will take place in the fall and post assessments will take place in the spring. To ensure an opportunity for growth, a minimum of 6 months will be designated between the pre- and post-assessments. All children will be assessed at their program. Researchers will schedule child assessments in advance and will request quiet space for assessment purposes. Child assessments will be conducted in English only. This is being done, in part, to provide a standardized protocol for all English language and dual language learners, since reliable and valid child assessment measures are only available in Spanish and English.[1] It is important to note that when examining outcomes, the study employs pre- and post-assessments of the same children which allows analyses to consider children’s growth, not just end-points and benchmarks. Thus, the diverse starting points that children from diverse language backgrounds will exhibit is taken into account, with an emphasis placed on gains (Zellman & Fiene, 2012). To ensure that English language and dual language learners are appropriate for testing, the initial set of the Peabody Picture Vocabulary Tests (PPVTs) will be given to all children to establish that they meet the basal requirements as specified by the PPVT. If children cannot meet these requirements, they will be considered ineligible for the assessment portion of the study and a different child will

[1] If an additional validation study were to be conducted with Family Child Care (FCC) educators, we would recommend assessment in both English and Spanish. This is due to the high number of FCC homes that are Spanish-only environments, potentially resulting in Spanish-related gains.

be randomly selected for assessment purposes. Data from these assessments will be used to answer research questions related to Approach 4.

Measures

Preschool Direct Child Assessment	**Domain**
The Peabody Picture Vocabulary Test-IV	Receptive language
Woodcock Johnson III Form A: Test of Achievement, Letter-Word Identification	Pre-literacy skills
Woodcock Johnson III Form A: Test of Achievement, Applied Mathematics	Early math skills
Preschool Teacher Rating Assessment	**Domain**
The Devereux Early Childhood Assessment Preschool Program, Second Edition (DECA)	Social emotional development, related to relationships and working with others
Preschool Learning Behaviors Scale (PLBS)	Academic-related social emotional competencies
Toddler Teacher Rating Assessment	
The Devereux Early Childhood Assessment for Toddlers	Protective factors and potential risks in the social and emotional development of toddlers

Children's receptive vocabulary will be assessed using the Peabody Picture Vocabulary Test-IV (PPVT-IV), designed to measure receptive vocabulary and verbal abilities (Dunn & Dunn, 2007). It is an individually administered, norm-referenced instrument which is scored based upon the age of the child. Scores on the PPVT-IV have been found to correlate with measures of general intelligence (Dunn & Dunn, 1997). The reliability and validity of the PPVT-IV are strong: internal consistencies range from 0.92 to 0.98 with a median of .95 and test-retest reliability ranges from 0.91 to 0.94. Correlations between the PPVT and other measures of verbal ability are: 0.91 (WISC-III VIQ), 0.89 (KAIT Crystallized IQ), and 0.81 (K-BIT Vocabulary).

The Woodcock Johnson III Form A: Test of Achievement, Letter-Word Identification & Applied Mathematics (WJ III) will be used to assess children's understanding of mathematical concepts and recognition of letters and words (Woodcock, McGrew, & Mather, 2001). It is an individually administered test and provides continuous age-based normative data obtained from information gathered on over 8,000 subjects in 100 geographically diverse communities in the US. The WJ III is a highly accurate and valid diagnostic system. Most of the WJ III tests show strong reliabilities of 0.80 or higher; several are 0.90 or higher and clusters show strong reliabilities, most at 0.90 or higher (Woodcock, McGrew, & Mather, 2001).

The Devereux Early Childhood Assessment Preschool Program, Second Edition (DECA) will be completed by classroom teachers for each of the randomly selected assessment children (LeBuffe & Naglieri, 1999). The DECA is designed to assess children social emotional development by asking teachers to evaluate the frequency of 27 positive behaviors (strengths) exhibited by preschoolers. These items were derived from the childhood resilience literature and through focus groups conducted with families and early childhood professionals. The DECA also contains a ten-item behavioral concerns screener. The assessment tool is designed for use with children ages three through five years old and is nationally standardized, reliable, valid, and easy to use.

The Preschool Learning Behaviors Scale (PLBS) will also be completed by the preschooler's teacher (McDermott, Green, Francis & Stott, 2000; McDermott, Leigh & Perry, 2002). It is used with children ages three to five and a half years and consists of 29 items, each focused on a distinct behavior related to learning, and is a measure of academic-related SEL competencies (Denham, Ji,

& Harme, 2010). It yields three learning behavior dimensions: competence/motivation (11 items), attention/persistence (9 items), and attitudes toward learning (7 items). Examples of items are: “Cooperates in group activities,” and “Is reluctant to tackle a new activity.” Inter-rater reliability coefficients rages from 0.57 to 0.73 and internal consistency ranged from 0.75 to 0.89 (McDermott, Leigh, & Perry, 2002; Schaefer, Shur, Macri-Summers, & MacDonald, 2004; McDermott, Rikoon, Waterman, & Fantuzzo, 2011).

The Devereux Early Childhood Assessment for Toddlers (DECA-I/T) will be completed by teachers in selected toddler classrooms (Powell, Mackrain, & LeBuffe, 2007). DECA-I/T is a tool for assessing protective factors and potential risks in the social and emotional development of toddlers. The assessment is a standardized, norm-referenced, reliable, and strength-based assessment, making it appropriate for evaluating toddler and program outcomes. The DECA toddler assessment has 36 items that reflect positive behaviors (strengths) typically seen in resilient toddlers. These positive behaviors comprise three protective factor scales: initiative, attachment/relationships, and self-regulation.

V. Sample Design

There are several issues related to selecting the sample for the validation study. Validation work linking ratings or rating components to children’s progress relies on accurately accounting for selection bias. Selection bias can occur at a number of stages in the QRIS. Programs participating in QRIS may not represent a random pool of programs in a state, limiting generalizability. Program numbers by region suggests that certain regions may be underrepresented in QRIS while other regions are over represented. Currently in Massachusetts, due to refinements and modifications in the system, there are very few programs at Level 4, which needs to be accounted for in the design. Finally, reviews of research demonstrate that the quality of early education and care programs is associated with higher language, academic, and social skills, and fewer behavior problems, but the effect sizes of these associations are small (Burchinal, Kainz, & Cai, 2011). It is important to put validation results in the context of this research and to note the efforts in the field to strengthen the breadth, depth, and content of the available quality measures (Zaslow, Marinez-Beck, Tout, & Halle, 2011).

The sample for the proposed study reflects the tradeoffs between budget and resources, number of programs at each level, and sample precision. The primary goal of the sample design is to provide reliable estimates that represent MA center-based programs in the QRIS. The challenge for the validation study is to develop a sample design that provides sufficient statistical power to meet its key goals, which are to assess (1) whether components of quality and the quality levels can be relied on as accurate indicators of program quality, and (2) the extent to which QRIS quality levels are associated with expected differences in children’s development and growth trajectories.

Program Sample

To meet the study objectives, we propose a multi-stage sampling design stratified by QRIS level. In the first stage, we will randomly select 120 center-based programs, stratified by QRIS Levels 1, 2, and 3. To ensure that we obtain a sample that is representative of each level, a pre-screen will be conducted at recruitment to eliminate programs from the sample that are likely to transition to the next level during data collection. The pre-screen will be a brief survey of QRIS transitioning activities to determine the likelihood of the program moving to the next level (e.g., actively working to finish training requirements with anticipated completion date within 3–4 months, waiting for QRIS program specialist visits). The elimination of programs likely to transition will allow the sample to be clearly defined by level, to remain stable, and to be stratified by level. We do, however, recognize that some programs may still transition during data collection. We have taken this in account when determining the needed sample size by level. Transitioning programs that are in the sample will be flagged and accounted for in analyses.

To ensure statewide representation, we will place a quota of a minimum of 12 programs from each of the five EEC regions for the total sample. Based on current numbers, we anticipate that the random sample will yield 120 center-based programs at Levels 1–3 serving preschool children; some of these programs will also serve infants and/or toddlers. It is important to note that the vast majority of center-based programs (approximately 98%) have at least one preschool classroom. However, only about 61% of center-based programs have an infant and/or toddler classroom. As such, the sample will be evenly stratified by level across preschool programs but may have uneven sample sizes for the infant and toddler programs by level. The recruitment process will be closely monitored for infants and toddlers to ensure adequate representation by level for the infant and toddler sample. If needed, additional infant and toddler programs will be added to allow adequate representation by level to support analyses.

We will also conduct a case study of programs at Level 4. At this point, Level 4 programs are being considered as case studies because of the uncertainty as to the number of programs that will be Level 4 certified at the time of data collection. We will randomly select up to 30 Level 4 programs for the validation study. There is however, a possibility that we will be unable to randomly select programs, due to the limited number of available programs. There is also a possibility that there will be significantly fewer than 30 programs at Level 4. If this does occur, we will collect data on all available Level 4 programs and will include these programs in the data analyses when statistically possible. If significantly fewer than 30 programs are certified at Level 4, other analyses will be conducted to describe these programs and their quality supports in order to provide insight into policy and practice.

Classroom Sample

One preschool classroom will be randomly selected by researchers from the selected programs, and, in programs serving infants and/or toddlers, one infant/toddler classroom. If programs serve only preschool-aged children, one preschool classroom will be randomly selected. If a program serves both preschool children and infants/toddlers, one preschool classroom with be randomly selected and one infant or toddler classroom will be randomly selected. Random selection of the infant and toddler classrooms will alternate by classroom type to achieve a balanced sample of 36 infant classrooms and 36 toddler classrooms at Levels 1–3 and an even number of infant and toddler classrooms at Level 4. Only one preschool classroom and one infant or toddler classroom will be randomly chosen from each of the selected programs to reduce the clustering of the sample. When we examine relations among preschool outcomes and quality indicators, clustering reduces precision and increases error because classrooms and children from the same program are more likely to be similar to one another as compared with those from different programs. Additionally, the selection of only one preschool classroom and one infant and toddler classroom reflects the QRIS model of the level being representative of the program, as a whole. As such, every classroom in the program should be in line with the level requirements.

Child Sample

The ultimate goal of the child assessments is to gather pre-and post-assessments of four preschool children per classroom, for a total sample of 480 children with pre- and post-assessments at Levels 1–3 and up to 120 children with pre- and post-assessments at Level 4. An oversample of two children per classroom will be employed in the fall to account for attrition and absences at the spring assessment period and allow sample goals to be met. To select the sample, all children in the randomly selected classroom will receive a study letter, along with a form to opt-out of the child assessment portion of the study. This passive consent processes has been approved through the New England IRB. On the day of assessment, trained child assessors randomly select six preschool children to be assessed (three boys and three girls). Assessments will take place in the morning at the child's school on a predetermined day. At the spring post-assessment period, four children from the original pool of six will be assessed. If five or six children are present at this testing point, only four children (two boys and two girls) will be randomly selected by the assessors for testing. Similar to the fall assessments, children will be tested at their preschool program on a morning predetermined in conjunction with the program and teachers.

For the toddler sample, we will ask teachers to complete a pre- and post-assessment of children's social-emotional development. Because there are fewer toddler classrooms, the study goal will be to have pre- and post-assessments of six children per classroom to achieve an adequate sample minimum of 216 children (not including Level 4). As such, we will randomly select eight children whose parents have consented, through the passive consent process, for the teacher to complete the DECA-I/T in the fall. In the spring, we will ask the teacher to complete a post-assessment on six children of the pool of eight. Similar to the preschool sample, the toddler sample includes an oversample of two children to account for attrition of children from the program and/or classroom.

Sample Selection	Level 1	Level 2	Level 3	Level 4	Total
Stage 1: **Programs**	40 programs	40 programs	40 programs	Case study of 30 programs	• 120 center-/school-based programs • Up to 30 additional Level 4 programs
Stage 2: **Classrooms**	40 preschool classrooms 24 infant or toddler classrooms	40 preschool classrooms 24 infant or toddler classrooms	40 preschool classrooms 24 infant or toddler classrooms	Maximum of 30 classrooms 20 infant or toddler classrooms	• 120 preschool classrooms at Levels 1–3 • Up to 30 preschool classrooms at Level 4 • 72 infant or toddler classrooms at Level 1–3 • Up to 20 infant and toddler programs at Level 4
Stage 3: **Children**	160 preschool children 72 toddlers	160 preschool children 72 toddlers	160 preschool children 72 toddlers	Up to 120 preschool children Up to 60 toddlers	• 480 preschool children at Levels 1–3 • Up to 120 preschool children at Level 4 • 216 toddlers at Levels 1–3 • Up to 60 toddlers at Level 4

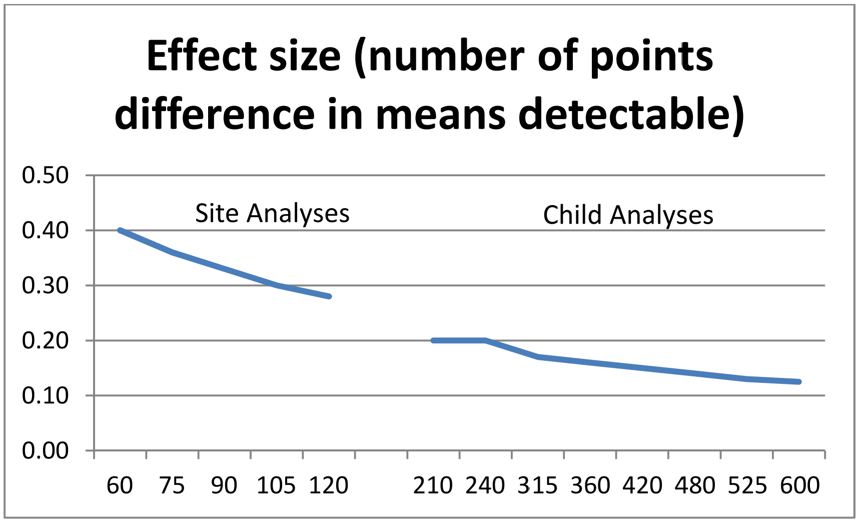
Sample Size and Statistical Power

In order to estimate total sample sizes for providers within each of the four quality strata, we must consider the minimum detectable effect size that is considered desirable. Effect size is defined as an estimate of the magnitude of the relationship or difference between two or more variables. Effect sizes of approximately 0.20 are considered "small," such as the difference in mean IQ between twins and non-twins (Cohen, 1988). Moderate effect sizes typically approximate 0.50 and describe differences large enough to be clearly visible in behavior or attitudes. Large effect sizes are those above 0.80 and are represented, for example, by the mean IQ difference between individuals with a graduate degree and those with a high school level of education. However, the delineation of thresholds for considering an effect size to be small, moderate, or large may vary depending on the research question being asked, and the methodological rigor of the study (Cohen, 1988).

In general, the smaller the minimum detectable effect size, the larger the sample required to provide sufficient statistical power for most of the key statistical analyses. Although one may want to choose an effect size that is as small as possible in order to have the ability to detect small differences, there are important budget and resource implications that might require compromises in sample precision. Thus, it is useful to consider a range of effect sizes and identify how the required numbers of providers will vary given each increase or decrease in the effect size.

To determine the meaningfulness of the minimum detectable effect sizes, it would also be useful to refer to the findings from past studies and to interpret effect sizes in terms of meaningful changes in key measures, such as the ECERS-R. The ECERS-R uses a seven-point scale and, based on other research and current QRIS data collection, we expect a standard deviation around 0.8. The proposed sample size for the study would allow for a mean difference of 0.28 points in the average total ECERS-R score or a subscale score, which is considered a small to moderate effect size. To be conservative, this does not consider the Level 4 programs. Inclusion of 30 Level 4 programs would allow for an effect size of 0.25 in average ECERS-R scores. The study proposes a smaller sample for the infant and toddler programs. The sample size of 72 allows for a mean difference across Levels 1–3 of about 0.4 in average total ITERS-R scores, which is considered a moderate effect size. Again, this does not consider the Level 4 programs, since the number is uncertain. Based on previous data collection for the QRIS, we expect greater variability in scores on the ITERS-R. Therefore, we believe that this sample will prove sufficient for the infant and toddler classrooms.

Since the Arnett can be used for all ages, the data collected for the preschool, infant, and toddler classrooms can be readily combined, resulting in a minimum sample size of 192 (Level 4 excluded), which would allow for small effect sizes to be detected. The chart below exhibits effect sizes by sample size for programs and children, using standards for the ECERS-R and the ITERS-R for the programs.



As the chart above indicates, a minimum sample size of 480 preschool children is sufficient to detect a small effect size of 0.14, using ANOVAs (analyses of variance) and regression analyses. To be conservative, this does not include the assessment of children at Level 4, which will increase the sample size and potentially allow a smaller effect size to be detected. Previous studies have documented small effect sizes regarding the association between quality and child outcomes. For the toddler assessment, a sample of 216 will be sufficient to detect a small to moderate effect size of 0.20, using ANOVAs and regression analyses, without including children at Level 4.

VI. Study Analyses

Approach 1. Examine the Validity of Key Underlying Concepts

1. *Do experts in the field in MA support the key components of quality in the MA QRIS? What are the perceptions of the QRIS and QRIS engagement by MA programs and providers?*

To address the first research question, highlights from the EDC report and findings from the pre-validation work done by UMDI through focus groups and a state-wide provider survey will be represented and synthesized. Both qualitative methods and quantitative descriptive data will be presented.

Approach 2. Examine the Measurement Strategies Used to Assess Quality

2. *What are the characteristics of programs by level? Are these characteristics consistent with QRIS level requirements?*

Observational and survey data for Levels 2, 3 and 4 will be compiled for each participating program to create a program QRIS profile. These profiles will be used to determine if the program meets the criteria for its current level. Rates of accuracy and false positives will be calculated to determine the reliability of the QRIS in classifying programs into distinct levels and the validity of QRIS assessment activities in determining the level achieved by programs. Data from Level 1 programs will also be analyzed to determine homogeneity of the group and identify potential program characteristics that may be related to the Level 1 status. Correlation analyses will be conducted to determine if specific characteristics of quality and the programs are related to QRIS levels.

3. *Do measures relate to one another as expected? Do different ways of calculating or combining scores yield more meaningful distinctions among programs?*

The overall question, in Approach 2, is whether each measure of quality, such as a QRIS indicator or a score from an observational tool, measures what it is intended to measure and whether it contributes uniquely to the overall level rating. For analyses, we will examine the distribution and variance of scores for a given indicator or set of indicators. If distributions are skewed or lack variance, then it is likely that the measures will not distinguish meaningful levels of quality. Additionally, we will conduct inter-correlations among the indicators. The strength of correlations between indicators will help to determine whether a given indicator is contributing unique information to measuring quality. If two indicators are highly correlated with each other, they are providing similar information and may not both be necessary in the rating. According to Zellman, and Fiene, indicators should be moderately correlated, indicating that they are related to each other but not redundant (2008). Additionally, the inter-correlations will provide important information regarding whether certain components are related in a manner that is expected. An exploratory factor analysis will be conducted to further explore how variables are grouped together and are related to form factors. The factors will be analyzed for the purpose of identifying significantly related variables and potentially reducing QRIS standards that are redundant. We will also conduct logistic regressions to determine which variables discriminate between level groupings. Logistic regressions are flexible in assumptions and types of data that can be analyzed. Logistic regression can handle both categorical and continuous variables, and the predictors do not have to be normally distributed, linearly related, or of equal variance within each group (Tabachnick and Fidell, 1996).

Approach 3. Assess the Outputs of the Rating Process

4. *Do programs in different QRIS levels differ significantly in observed and structural quality? Do level distributions vary by key program characteristics?*

Analyses will employ ANOVAs to compare the overall quality scores on the ECERS-R/ITERS-R, the individual subscale scores on the ECERS-R/ITERS-R, and the subscale scores on the ARNETT by QRIS level to determine if significant differences are found in observational quality by programs' QRIS level. Additionally, ANOVAs will be conducted to determine differences in key quality indicators related to the five quality domains of the QRIS. We will also conduct analyses to determine if key program characteristics (e.g., multi-service organizations versus care-only organizations) are related to level designation.

Approach 4. Examine How Ratings are Associated with Children's Development

5. *Do children who attend higher-rated programs have greater gains than children who attend lower-level programs?*

It should be noted that studies examining relationships between outcomes and quality indicators have found only modest effect sizes. Additionally some studies from other states have not be able to substantiate a significant relationship between QRIS levels and child outcomes (Zellman & Fiene, 2012). The proposed study has incorporated a statistically large enough sample size to allow researchers to detect moderate to small effect sizes. Additionally, the analyses explore relations between QRIS levels as well as key QRIS indicators. Lastly, the study incorporates pre and post data from the same children to assess development **growth**, as opposed to developmental end points at the conclusion of the preschool year. This will allow the study to account for different starting points on measures of key developmental domains, such as language and literacy, early math, and social-emotional development.

Correlations will be conducted among QRIS levels and the various measures of development growth as determined from differences in pre- and post-assessments. Next analyses will include an ANOVA, testing group differences by QRIS level across the various measures of child outcomes. The use of ANOVA in validation research is a common (e.g., McKenzie, Stone, Feldman, Epping, Yang, Strikmiller, Lytle, & Parcel, 2001; Penny, Creed-Kanashiro, Robert, Narro, Caulfield, & Black, 2005; Schnitzer, Andries, & Lebeer, 2007) and useful way to examine group differences. One of the assumptions of an ANOVA is the independence of observations. Given that the children are clustered within classrooms, analyses will only include the program characteristics of the QRIS level, as opposed to individual classroom level variables (e.g., ECERS-R scores). This avoids unmeasured classroom effects that can make children within a classroom more similar to one another than to children in other classrooms. Additionally, an exploratory factor analysis will be conducted to determine if the developmental growth measure found across the various child assessment tools forms a latent variable of overall developmental growth. If a latent variable is confirmed, an ANOVA will be conducted to determine group differences by QRIS level in overall developmental growth of children.

6. *What QRIS standards are significantly associated with increased child outcomes? Does QRIS level significantly predict children's outcomes beyond key demographics of children?*

Multiple regression analyses will be performed to determine if the QRIS level predicts outcomes above and beyond key child demographics, including subsidy eligibility and home language. The analyses will also explore relations among key QRIS standards and child outcomes. Correlations will be conducted to explore the relations among specific aspects of QRIS indicators and child outcomes. Regression analyses will also be performed to determine the specific aspects of the QRIS level rating that are particularly salient in predicting children's outcomes. To stay consistent with the QRIS as an assessment of program level, these analyses will incorporate program-level quality indicators as opposed to classroom-level variables to avoid clustering the sample. If power and variability is sufficient, exploratory HLM analyses will be performed using both child and classroom level co-variants to analyze associations among key classroom-level indicators and outcomes.

VII. Timeline

The following timeline is proposed to meet study tasks and deliverables.

Scope of Services: January 1, 2014 – June 30, 2014

1. Statewide QRIS Provider Survey – finalize and report
2. Validation Study Redesign – meet with study advisory board, EEC commissioner, EEC policy subcommittee
3. Prepare for Fall Data Collection, Sample Selection, and Program Engagement
4. Study Sample Selection and Engagement

Scope of Services: July 1, 2014 – June 30, 2015

1. Prepare for Fall 2014 Data Collection
2. Fall 2014 Data Collection – child assessments
3. Winter/Spring 2014-15 Data Collection – ERS observations, program information from directors and teachers
4. Spring 2015 Data Collection – child assessments

Scope of Services: July 1, 2015 – December 31, 2015

The focus of work during the final six months of the QRIS Validation Study will be to conduct data analysis and generate findings in accordance with the study's key research questions to be prepared for EEC and key stakeholders. Findings of each data collection phase will be integrated into a final project report that offers an overarching view of study findings and provides salient suggestions for QRIS improvements accumulated over the course of the study.

VIII. Assumptions

There are several assumptions built into the design of the study. Without these assumptions being met, the study design would need to be altered. They include the following:

- EEC will have verified a sufficient number of programs at Levels 1, 2 and 3 by mid-June 2014.
- EEC will encourage program participation in the validation study, including all incentives for provider's participation in the study.
- Outside of the terms of the contract with UMDI to conduct the study, EEC will fund the cost of 80 ERS assessments as part of the study, to include programs at any QRIS level.
- EEC will work with researchers to develop the recruitment prescreen survey to ensure transitioning activities are accurately identified.
- EEC will participate in regular meetings with researchers to inform study implementation and data collection.
- EEC will provide UMDI with timely notification of any planned changes to QRIS that may have implications/impact of the validation study design.
- EEC will provide timely feedback on the draft final report to facilitate completion of data analyses and the final policy and practice recommendations.

IX. Deliverables

The study design includes the following deliverables:

- Final validation study research plan submitted to EEC by March 3, 2014, for approval by EEC.
- Copy of study approval letter from the Institutional Review Board (IRB) by April 30, 2014.
- Final instruments for program quality measures shared with EEC by April 30, 2014.
- List of at-scale providers selected for study along with summary statistics, including program characteristics (e.g., program size, level, location, etc.) by June 30, 2014.
- Upon completion of all Summer 2014 trainings required for study data collection (e.g., direct child assessments) UMDI will provide EEC with a list of attendees and results of the training, including successes in achieving reliable raters, by September 30, 2014.
- Teacher rating materials including rating forms, directions, and answer sheets created by UMDI for ease of administration and data quality will be shared with EEC by September 30, 2014.
- Upon conclusion of initial child assessments and teacher rating scales, UMDI will provide EEC with a summary of administration results by January 30, 2015.
- Upon completion of all Winter/Spring 2014-15 trainings required for study data collection (e.g., Environment Rating Scales) UMDI will provide EEC with a list of attendees and the results of the training, including successes in achieving reliable raters by March 3, 2015.
- Upon conclusion of second child assessments and teacher rating scales, UMDI will provide EEC with a summary of administration results by June 30, 2015.
- UMDI will provide EEC with a draft report including summary statistics of all program quality measures by June 30, 2015.

- UMDI will prepare a QRIS descriptive report containing a complete summary of participating providers within QRIS and how they fared on the varied data collection measures used in the study by November 15, 2015.
- UMDI will prepare and submit a draft final report of validation study findings for discussion with the QRIS Validation Study Advisory Board and with EEC. Based on these discussions, UMDI will revise the draft of study findings and produce a final report for formal presentation and approval to the EEC Board by December 31, 2015.