EFFECTIVE EVALUATION OF QUALITY RATING AND IMPROVEMENT SYSTEMS FOR EARLY CARE AND EDUCATION AND SCHOOL-AGE CARE

Research-to-Policy, Research-to-Practice Brief OPRE 2011-11a
June 2011

OPRE
Office of Planning, Research and Evaluation
Administration for Children and Families

Disclaimer:
The views expressed in this publication do not necessarily represent the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families or the U.S. Department of Health and Human Services.

Acknowledgements
The authors would like to thank Ivelisse Martinez-Beck, Kathleen Dwyer, Mary Bruce Webb and Naomi Goldstein at the Office of Planning, Research and Evaluation and Kathryn Tout at Child Trends for their guidance and feedback on this paper.

Effective Evaluation of Quality Rating and Improvement Systems for Early Care and Education and School-Age Care

Research-to-Policy, Research-to-Practice Brief OPRE 2011-11a
June 2011

Submitted to:
Ivelisse Martinez-Beck, PhD., Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Submitted by:
Gail Zellman, RAND Corporation
Richard Brandon, University of Washington, Human Services Policy Center
Kimberly Boller, Mathematica Policy Research
J. Lee Kreader, National Center for Children in Poverty

Contract Number: GS10F0030R
Project Director: Kathryn Tout
Child Trends
4301 Connecticut Ave NW
Washington DC, 20008

Suggested Citation:
Zellman, G. L., Brandon, R. N., Boller, K., & Kreader, J. L. (2011). *Effective Evaluation of Quality Rating and Improvement Systems for Early Care and Education and School-Age Care,* Research-to-Policy, Research-to-Practice Brief OPRE 2011-11a. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This Brief was developed by members of the Quality Initiatives Research and Evaluation Consortium (INQUIRE) which is designed to facilitate the identification of issues and the development and exchange of information and resources related to research and evaluation of quality rating and improvement systems (QRIS) and other quality initiatives. INQUIRE is funded by the Office of Planning, Research and Evaluation through the Child Care and Early Education Policy and Research Analysis and Technical Expertise contract with Child Trends.

Effective Evaluation of Quality Rating and Improvement Systems for Early Care and Education and School-Age Care

Goals of this Brief

It is important to evaluate Quality Rating and Improvement Systems (QRISs) so that policy makers and stakeholders can learn how well they are working and how they might be improved. Well-designed QRIS evaluations go beyond a "pass/fail" judgment to identify implementation successes and problems and assess what needs to be done to improve the system. A recent evaluation of a small QRIS field test in two communities in Washington State exemplifies what good evaluations can do. The evaluation provided useful findings for both policy makers and program developers by examining both "how" the QRIS was working (an implementation study), and whether the improvement approach was affecting child care quality.[1] The findings informed a larger field test in five locations around the state.

A QRIS is a market-based quality improvement initiative intended to improve program quality and child outcomes by promulgating standards, rating providers on their attainment of those standards, and publicizing those ratings. Quality improvements are encouraged through a range of incentives and supports. By using published ratings to select higher-rated providers, parents also encourage providers to obtain higher ratings.

The purpose of this document is to encourage high-quality QRIS evaluations by providing timely information on evaluation options to those who may be in positions to authorize, finance, design, and refine QRISs and other quality improvement efforts, including state child care administrators, legislators, and other potential funders such as foundation personnel, as well as child care and early education provider organizations.

This Brief presents basic evaluation concepts, useful tools for determining the appropriate design and timing of an evaluation, and evaluation references and resources for those who wish to learn more. Readers should come away with enough background to understand the basic issues in designing evaluations of quality improvement efforts and a stronger sense of their importance.

Reasons to evaluate

The overriding motivation for evaluating any policy or system is to find out if and how it "works" and how it might be improved. "Is it working" may include a wide range of issues, addressed through questions and responses that range from "are all the pieces in place?" and "is it achieving its goals?" to "does it produce better outcomes than another policy with the same goals or than business as usual?", and "given what it does, is it worth the cost?" Since QRISs aim to improve quality based on a market-based logic model, a good evaluation can also test that underlying model and assess whether the model represents an effective way to improve the early childhood education and school-aged care (ECE-SAC) system. Depending on how the evaluation questions are posed, an evaluation often can produce far more than a "yes/no" or "works/doesn't work" answer. A well-designed evaluation can help to pinpoint problems with design, implementation or funding that need to be corrected before it is reasonable to assess whether the QRIS is achieving its goals. These problems may include inadequate supports or incentives, lack of provider or parent understanding of the system, or weaknesses in the measures used to assess key outcomes. It can also assess whether there are equally positive impacts for sub-groups of children, providers or communities.

Why evaluate a QRIS?

- To inform program design, management, improvement, and scale up
- To address accountability requirements and assess efficiency
- To support comparisons of outcomes across different types of early childhood program investments

Exploring whether a policy works is important because the implementation of any policy imposes a series of financial and other costs; policymakers want to make sure that these costs produce commensurate benefits. In addition, the implementation of a given policy often precludes the adoption or implementation of other policy options; if a selected policy is not meeting its goals it may be necessary to modify the policy or the goals, or replace the policy with another approach.

Evaluation, done well, can also help to clarify the reasons behind any changes that appear to be attributable to the policy; if there is apparent progress, a strong evaluation design will enable policymakers to determine whether that progress is in fact a result of the policy or is an artifact of other factors. Careful evaluations of QRISs also will help policymakers to determine if the costs of these systems are producing commensurate benefits. Finally, a well-designed evaluation can help to identify unintended consequences such as frequent staff moves to improve ratios. Ultimately, a good evaluation should test the concept by answering the question: is this initiative meeting its goals in a cost-effective way? If not, it may be necessary to go back to the drawing board and rethink those goals, the nature and adequacy of the incentives and supports built into the underlying logic model, and the policy paths not taken.

Establishing a QRIS entails recognition that current public and private structures are not producing sufficiently high quality early learning opportunities, nor making them financially and geographically accessible to all children. Those who promote adoption of QRISs therefore have a stake in evaluating them, both to learn whether the factors motivating the QRIS are being remedied, and also to learn how the strategy and tactics being employed can be improved.

Factors that may impede support for evaluation

Despite these powerful reasons to carefully evaluate QRISs, there are also substantial barriers. Four important ones are discussed briefly below; each is followed by suggestions about how to respond.

Betting on Success. In order to marshal the necessary political support to adopt a new policy, a strong argument must be made for its value. This is particularly true for an ambitious policy such as a QRIS that may require substantial resources and often will subject previously unmeasured processes to public scrutiny. To line up the necessary political support, it may be necessary to assert that the policy *will* result in the achievement of the goals it is designed to address, rather than presenting the policy as a logical approach, a good bet, or a best-available option. In the process of winning support, there may be little room for expressing doubt or uncertainty.

When the adoption of a QRIS represents the successful conclusion of a contentious struggle or the seizing of a rare opportunity, the decision-makers and advocates may come to associate "victory" with the mere creation of the QRIS, leaving little enthusiasm for evaluating its long-term impact. Evaluations always carry risks; a positive outcome is never certain. Under such circumstances, the tendency to "declare victory now" is strong. A related tendency is to conduct an evaluation limited to collecting participants' views; such "evaluations" often produce laudatory results that are useless for accountability or program improvement purposes.

- ✓ **How to Respond**. Efforts that are not evaluated and cannot demonstrate impact are on shaky ground for continued policy and budgetary support. If it becomes clear that the conditions leading to adoption of the QRIS persist—poor child outcomes, low quality ECE-SAC—then the entire effort may be in jeopardy. Healthy skepticism is an appropriate and productive stance to take toward any public policy.

Going for All or Nothing. Politicians and advocates who have put their prestige and power behind a new policy may not want to learn that it is not working, even if an evaluation might help to pinpoint places where improvements might change the policy's trajectory and improve its ultimate outcomes. While researchers might argue that it is important and beneficial to study the implementation of a policy and discover that its effect was neutral or even negative, policymakers don't always share that view.

- ✓ **How to Respond**. Evaluating a policy like a QRIS conveys an understanding that the current approach is but one of many possible permutations; a particular configuration was selected because it seemed the most efficient or productive approach within current constraints, but there could have been others. Framing an evaluation in this way is important because if the evaluation shows that the adopted approach is not working well, there will be less inclination to reject the entire idea out of hand and more willingness to reexamine which aspects were more and less effective in an effort to refine the policy and/or its implementation.

Comfort with the Status Quo. Evaluation may also be resisted when evaluation findings might undermine a carefully negotiated division of power or resources or call into question an approach that garners more political support than others. When service providers become comfortable with an existing resource distribution formula or set of standards, their interest in evaluating the degree to which the funds are effectively addressing the issues they were designed to address may wane. Assessing the value and use of a QRIS might force allocation decisions to be reopened and challenging distribution and rating process decisions to be restarted. In particular, if the evaluation shows that the level of incentives and supports provided are insufficient to produce desired changes, it can create pressure for substantial funding increases that policy makers may not want to address.

- ✓ **How to Respond**. Better outcomes for children represent the driving force behind QRISs. With this sort of resistance, it may be necessary to assert the importance of children's needs over provider or policymaker concerns. It can be pointed out that the incentives and supports of a QRIS address providers' needs. It can also be useful to point out that the world is not standing still: the demands on school-age children are increasing as the global economy becomes more competitive. In this context, a strong early learning foundation becomes more essential each year.

Research Challenges. A potentially important reason why there are few evaluations of QRISs is that such research is challenging. For example, to study the effects of a child care QRIS on kindergarten readiness, children must be followed over time and receive care in a given program during the study period. If children change providers during the study or leave formal child care entirely, their outcomes are no longer a fair test of the effects of a given provider on children's outcomes. Yet high attrition rates are common in child care settings as families move, parents lose and change jobs, and children age out of care. In addition, understanding incentives and price effects requires obtaining financial data from providers, which can be difficult and costly. Policy makers often face "sticker shock" at the cost of a well-designed and well-implemented evaluation. But the cost of continuing an ineffective policy is far greater. As long as evaluation does not inform policy, opportunities to improve policy or replace it with more effective approaches are lost.

- ✓ **How to Respond**. The cost of evaluation is small compared to the total public expenditures for ECE-SAC—far less than 1 percent of a state's CCDF or public pre-K funding, and an even smaller percentage of what parents privately pay for ECE-SAC of uncertain quality. QRIS funders and developers can make a number of early decisions that would make evaluation less costly and more likely to find effects. For example, before a QRIS is implemented, baseline data on quality of ECE-SAC and child well-being could be collected. Quality rating data collection can be designed to serve evaluation purposes, avoiding the need for additional data collection, an expensive evaluation component. Plans to phase in the system over time or in different geographic areas allow powerful research designs to address accountability and program improvement needs. When policy makers, QRIS developers, and evaluators collaborate early in the development process, relatively simple and low-cost design elements that set the stage for rigorous evaluations can be built into the system.

Evaluation is feasible and essential

Effective evaluation of QRISs is both necessary and feasible. QRISs represent a primary strategy states are employing to improve ECE-SAC quality. It is imperative that we learn whether we are on the right track. The feasibility of evaluation has been demonstrated in several states and localities. In none of these states has the evaluation resulted in declining public support or provider participation; rather, it has led to redoubled efforts to refine the strategy. This brief discusses practical ways to address evaluation challenges and conduct meaningful evaluations with limited resources.

Logic models: The essential starting point for evaluating QRISs

A key tool for understanding QRISs and considering how best to evaluate them is a logic model. Logic models describe the process that ideally underlies the development and successful implementation of any new program or policy. Basically, a logic model is a systematic and visual way to present the relationships that are expected to exist among the *resources available* to the effort or program, the *activities or polices* that are to be put in place, and the *changes or results* that are expected to follow.[8] Logic models provide stakeholders with a road map describing the sequence of related events connecting the need for the planned program or policy with the program's desired results.

A well-articulated logic model describes key steps in the process and key outputs of each step. Those outputs, when well-defined, identify measurable behaviors or indicators at each stage of the implementation process, e.g., "meetings are held at least quarterly between specified actors" rather than "more collaboration occurs."[4,7] These indicators constitute the measures of the initiative's progress toward meeting its stated goals and should comprise key components of any evaluation design.

Logic models link the purpose, goals, objectives, and tasks included in a policy design with activities and expected outcomes and link back into program planning and resource allocation.

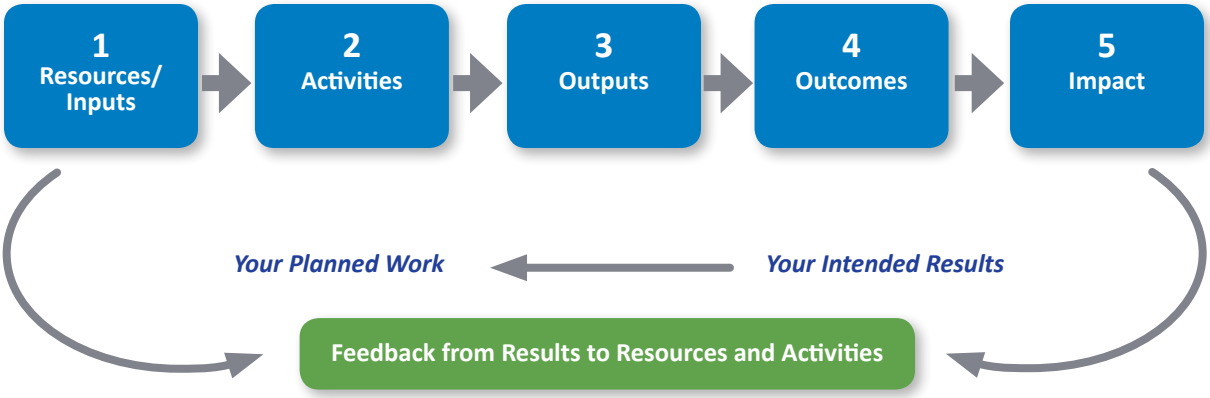
Presented graphically (see examples below), logic models display designers' theory of how proposed activities and policies will lead to desired goals through a logical chain of "if-then" relationships. Each component of the logic model should be accompanied by appropriate *measurable performance indicators* that are tailored to the activities specified. Performance indicators should be *specific* (understandable to users; easy to tell apart), *measurable* (data can be collected within constraints of time, cost and confidentiality), *unique* (select best one or two indicators instead of long list) and *robust* (not subject to manipulation to make performance look better).[6] The logic model should also make clear how long it is expected to take to achieve each specified change in behavior or conditions.

A logic model also helps planners and stakeholders to understand the developmental phase of a given policy and helps to clarify what should be measured to gauge progress at different points in the implementation process. Most policies require a considerable amount of time to be fully implemented; QRISs are no exception. By laying out the steps in an implementation process and specifying outputs and performance measures for each step, a logic model can pinpoint how far a policy has come in reaching its ultimate goals. This can help to manage expectations, e.g., it is unreasonable to expect improved child outcomes when parents don't yet know about the QRIS, and target evaluation efforts and output indicators to the appropriate level in the life of a new policy or program.

Several examples of logic models are presented below. The first is generic and simple, illustrating the overall structure of any logic model (Figure 1). The others are QRIS-specific, showing how that structure can be applied to a specific policy tool. The Zellman and Perlman model (Figure 2) focuses on parents and providers and articulates the process assumed to be involved in implementing a QRIS in some detail. In particular, it indicates the multiplicity of changes in behavior that are required to achieve the desired outcomes. The Brandon model (Figure 3) focuses more attention on financial assistance to both providers and parents, which emphasizes the role of financial supports and constraints in a QRIS. It also includes an additional QRIS outcome: reduction of gaps in child outcomes across children with different characteristics. Together, they illustrate how a logic model is developed and how it can guide both implementation and evaluation design.

The Basic Logic Model from the Kellogg Foundation Logic Model Development Guide (Figure 1) indicates the key categories that should be included and the order in which they occur. It clearly indicates a flow from Resources to Impact; we have adapted the Kellogg model to show the important feedback loop from outcomes and impacts to resources and activities. Understanding the relationships between outcomes, outputs, and activities can feed back into improved design of activities and more willingness to invest in the effort.

Figure 1 – Kellogg Foundation Basic Logic Model [Adapted to show feedback]



This model is read "left to right," emphasizing how the currently available resources and activities lead to measured outputs and outcomes. However, for strategic planning, it is often most useful to read a logic model from "right to left," starting with the impacts and outcomes desired, then considering what activities are likely to yield those outcomes and what resources are required to implement those activities.[3] The same principle can be applied to the Zellman and Perlman model (Figure 2) which is presented with the inputs at the bottom and the ultimate outcomes at the top.

Figure 2 – QRIS Logic Model[9]

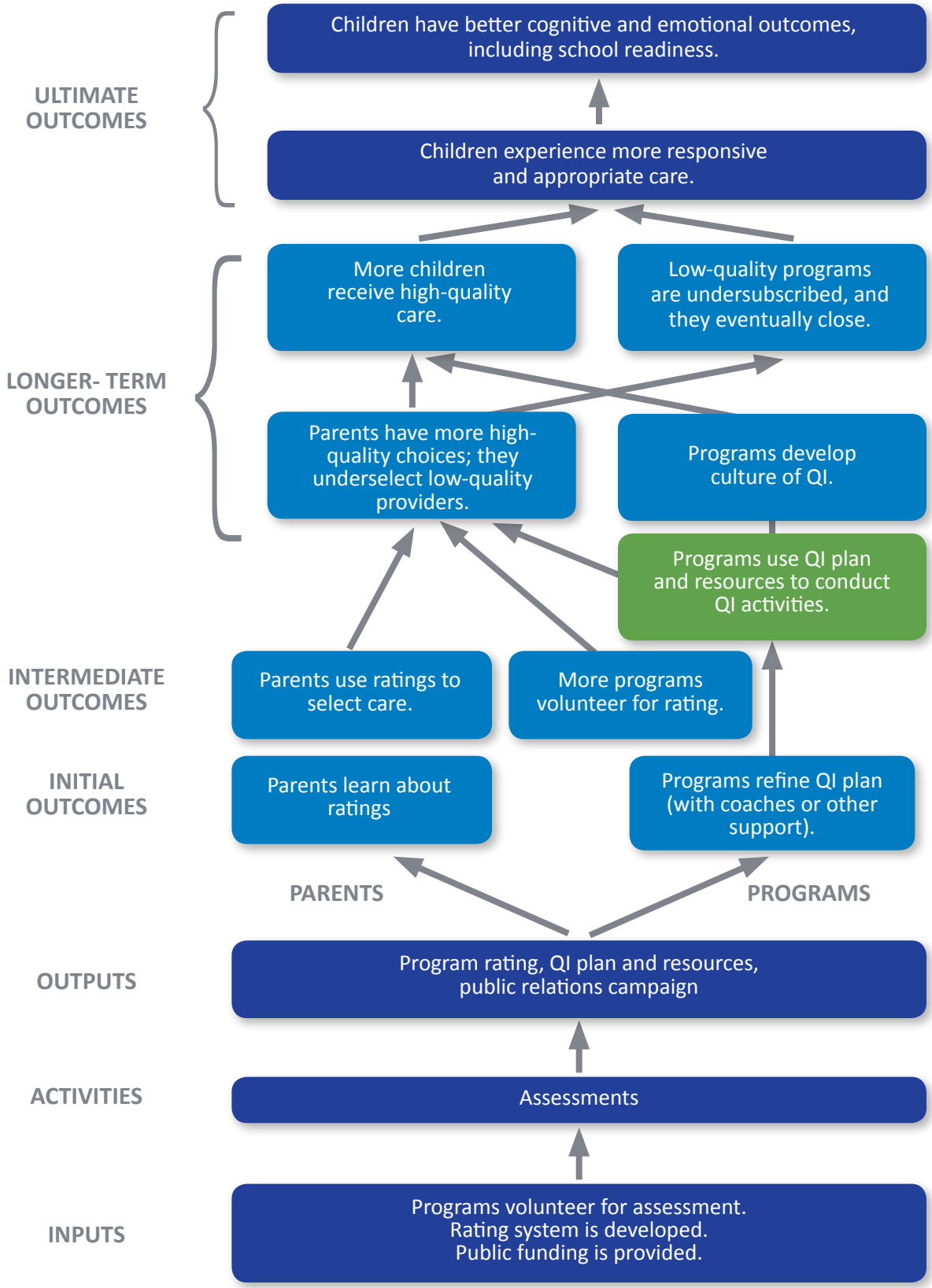
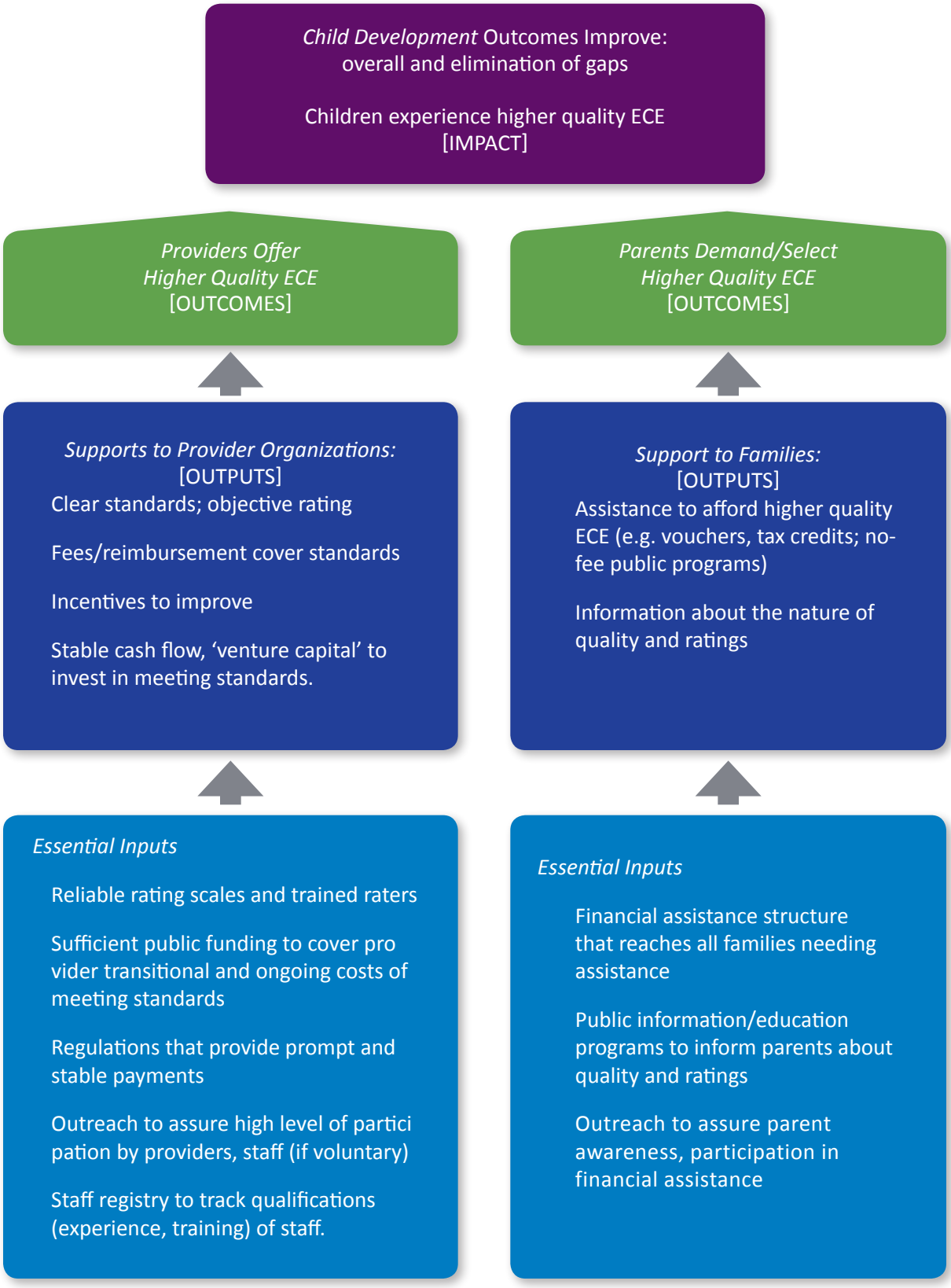


Figure 3 – Simplified Logic Model[2]



Testing logic models: Key research questions

A comprehensive logic model will suggest a variety of research questions that might be addressed in an evaluation or set of evaluations. However, it might not be feasible to address all of them, or to address all of them at once. For example, all QRISs aim to improve ECE-SAC quality with the expectation that this will ultimately improve child outcomes. However, directly assessing the effects of a QRIS on child outcomes is expensive, since it requires periodic assessments of a large number of young children and the need to obtain parental consent. If resources are not available, it may be necessary and wise to focus limited resources on measuring observed quality of adult-child interactions, relying on the research in the field that links such quality to child outcomes to infer effectiveness.[5] In this section we present a comprehensive set of topics that would contribute to understanding all aspects of a QRIS, knowing that in most cases not all of them can be included within a given evaluation design. However, one of the central concepts of a logic model is that each component is directly linked to those that precede and follow it; when selecting components for an evaluation focused on short-term outcomes such as ECE-SAC quality or parent awareness and use of QRIS to guide ECE-SAC choices, it is important not to skip any earlier links in the chain.

Systems and Markets:

A *system* is the stable pattern of interaction that may be planned or may emerge over time among many individuals and institutions. Within a system, actors have distinct roles, values, expectations, and resources. Typically, systems contain features that serve to rein force a stable pattern of interactions, but they can and do respond to planned interventions.

A *market* is a set of transactions between (in the case of QRISs) buyers (parents) and sellers (providers) of a good or service (ECE) at prices determined by the perceived value of the service and the resources and constraints of buyers. The quantity and quality of goods or services offered and purchased are influenced by the interaction of buyers' preferences or demands and sellers' ability to understand and respond to those demands.

One of the special attributes of a QRIS is that it is a market or system-level initiative, intended to affect not just the individual providers, most of whom volunteer to participate and the children they serve, but to have far broader, systemic impacts. By establishing quality standards, making participating provider quality widely known, and providing quality improvement incentives, the entire population of a jurisdiction (state, county, municipality, or school district) may be affected as parents begin to understand quality, demand it in their providers, and providers see that quality improvements are doable and likely to pay off. Both the Zellman et al. (Figure 2) and the Brandon (Figure 3) logic models reflect these effects, showing longer-term outcomes as the *overall range* of quality of ECE-SAC offered by providers and selected by parents, not just the quality of participating providers or programs. The Zellman model details the steps required for this to happen, with parents "under-selecting" low-rated providers, leading to their closing, and presumably moving to higher-rated providers, increasing their share of children enrolled. The Brandon model indicates that parents can either select different providers or demand that current providers improve. These choices affect the "market": parents exercising their role as consumers—informed by quality ratings and provided financial support that enables them to make choices—influence which providers flourish, improve, or wither. Providers—supported by professional development and responding to financial incentives—improve the quality of the services they offer. Thus, the introduction of information, supports, and incentives changes the nature of the transactions in the ECE-SAC market. When information, supports, and incentives are made available to all providers and parents in a jurisdiction (state, county, municipality, school district), then the resulting changes in both parent and provider behavior can transform the ECE system in that jurisdiction.

Consistent with our suggestion of backward mapping from outcomes, we begin our focus on testing logic models by specifying research questions that assess these market- and system-level effects, then work back to test the logic model components that are more focused on individual providers, staff or parents. There are other potential system-level effects on ECE-SAC quality that are not necessarily integral to the concept of a QRIS, such as improved collaboration and coordination of services. We have not included them in our discussion, although if a particular QRIS includes these system-level outcomes within its logic model and program design, they could be added to the list of evaluation questions.

Time is a critical dimension when considering which components of a logic model to include in a QRIS evaluation. As noted above, full implementation of a QRIS often takes some time, and may be an iterative process that relies on the outcomes of targeted pilots. When overall ECE-SAC system change is sought in a state or community, the supports and incentives that the QRIS provides must remain in place for an extended period before such system-level effects can be expected to occur. Evaluation designs must be sensitive to this reality and focus on those outputs and outcomes that can reasonably be expected to demonstrate changes in the time frame of the evaluation. Doing more—e.g., assessing effects on child outcomes in the first year of a QRIS, before improvements in earlier outcomes such as provider quality are demonstrated—could waste resources and virtually guarantee unnecessarily bleak results. For example, the measurement of impacts on children in an experimental evaluation, discussed below, may make much more sense after short-term impacts such as changes in ECE-SAC quality or provider training have been established. However, it is important to build a plan to do this over time into the evaluation design at the beginning, since many will want to know that such outcomes will be addressed at some point.

Below, we lay out some research questions derived from Zellman et al.'s and Brandon's logic models that are likely to be of interest to policy-makers and stakeholders. Research questions can be posed in two forms: (A) are objectives being met, and (B) what can be learned about the methods or situations affecting whether they are being met. The first set of questions focuses more on the outputs and outcomes of each part of the logic model; the second more on what explains whether each part leads to the next ones as anticipated. We also ask questions about the distribution of demonstrated effects, e.g., what is happening to *average* levels of intended outcomes, how many children and providers are at the top and bottom levels of quality and development, and whether gaps among different groups of children, parents, or providers persist. These important questions are illustrative; many more could be added to fully understand what is happening with a QRIS. Since QRIS is a market-based intervention, we start with the broader market- and system-level impacts and work back through the various outcomes (changes in parent and provider behavior) and inputs/outputs (supports and incentives offered) that are expected to precede and intended to produce those changes.

1. *System level effects:*

This set of questions addresses the impact of a QRIS on the ECE-SAC system and its use across an entire jurisdiction or community. There is emphasis on both the overall level of impact, and whether the impact varies for sub-groups of children, parents, or providers.

(A) Meeting objectives (outputs and outcomes)

- Does the overall level of development (cognitive, social-emotional, self-regulatory) among children residing in the market or jurisdiction improve over time; do gaps in development among advantaged and disadvantaged groups decrease or persist?
- Does the average level of quality offered by providers increase; does the percentage of providers offering ECE-SAC that does not meet minimum standards decrease toward zero; do a higher percentage of providers attain the top 1-2 levels of quality?
- Do gaps in the average quality experienced by children from advantaged and disadvantaged backgrounds, or different race-ethnic groups decrease?
- Does the overall level of staff quality, defined by their qualifications and their interactions with children and parents increase; do average levels of compensation increase to allow recruitment and retention of well-qualified staff?

(B) Factors affecting whether system-level objectives are being met

- What is the relationship between the type and quality of providers and levels of child outcomes, controlling for social-economic differences; for which disadvantaged groups are gaps being closed, for which do they persist?
- What attributes of providers are related to whether their quality improves, e.g., staff qualifications, participation in professional development opportunities, prices charged or rates of reimbursement received, management practices?
- What practices are related to improved staff interaction with children and parents: level of compensation; use of performance-based pay; management practices; participation in professional development; enhanced formal education or other qualifications; provider price/reimbursement levels? Do the racial-ethnic-linguistic backgrounds of staff roughly match that of the children for whom they are responsible and are such matches related to child outcomes?

2. *Changes in Provider Behavior: Organizations and Individual Staff*

This section poses questions about the degree to which changes in behavior by provider organizations and individual staff that are purported to lead to improved quality occur as anticipated. Questions about the factors affecting the degree of provider and staff changes, and whether changes are different among sub-groups or providers are then posed. These questions focus on whether the information, supports, and incentives offered in the QRIS are effective.

(A) Meeting objectives (outputs and outcomes):

- What percentage of provider organizations participate in voluntary QRIS? Are participation levels increasing over time?
- Do provider organizations continue to upgrade their quality toward the highest level, or do they plateau at some lower point?
- What percentage of individual staff participate in and complete voluntary opportunities to improve competence and qualifications? Which opportunities are related to actual increases in observed staff performance?

(B) Factors affecting whether objectives are being met

- What factors predict QRIS participation: neighborhood characteristics; socioeconomic status (SES) of local parents; management practices; what barriers to participation do providers report? Do providers report that they consider ratings fair and accurate; which components of the rating system do they consider better or worse?
- What quality improvement supports do providers select from among the available options: coaching and mentoring; transitional grants; professional development (PD) opportunities for their staff; higher reimbursement rates or quality bonus payments? What do providers report is/not appealing about these different opportunities?
- Does participation in different types of support predict the level of quality actually attained?
- What characteristics differentiate providers who keep improving toward the maximum level from those who plateau at lower levels?
- What percentage of individual staff participate in and complete voluntary opportunities to improve competence and qualifications: college or community-based courses; workshops; performance-based pay; listing on a registry? Which of these opportunities are related to actual increases in observed staff performance; what barriers do staff report to participating in these opportunities? What is the frequency and duration of these PD opportunities?

3. Changes in Parent Behavior

The research questions posed in this section address the degree to which parents are utilizing and responding to the information, supports and incentives they are offered through the QRIS to encourage the enrollment of a higher percentage of children in higher-rated settings. They attempt to consider a variety of potential barriers and determine whether the design and implementation of the QRIS are likely to help planners avoid them.

(A) Meeting objectives (outputs and outcomes):

- What percent of parents change provider or demand improvements from their current provider?
- What level of awareness and knowledge of the quality rating system do parents have?

(B) Factors affecting whether objectives are being met

- What factors affect parents' choice of provider? What barriers to enrolling at a higher quality provider do they report, e.g., cost; location; participation in public programs, assumptions about the link between cost and quality?
- What parent characteristics affect their level of QRIS knowledge and awareness?

4. Adequacy and Effectiveness of Inputs (Human and Financial Resources)

The questions in this section address the adequacy, equity and value of resources provided to the various participants in the QRIS, asking whether the resources are sufficient to implement the planned activities in a manner that will produce the desired changes in behavior.

(A) Meeting objectives (inputs and outputs):

- Are the inputs of sufficient scope to have the desired impact?
- Are quality ratings based on scales and methods that have been validated? Are they conducted in a timely, thorough manner? If not, why not? Are sufficient staff or contractors deployed to conduct timely ratings?
- Do most parents in the jurisdiction receive sufficient information regarding quality, price and other factors to exercise well-informed choice?
- Are parent information campaigns conducted in a thorough manner, using multiple media, languages, and trusted messengers?

(B) Understanding the relationship of inputs and outputs :

- Do all providers in the jurisdiction have access to the supports and financial incentives that are expected to improve quality; is support available for sufficient time to achieve the objectives?
- Do higher reimbursement rates or quality bonus payments cover providers' full cost of meeting higher quality standards; are all families who cannot afford the cost of higher quality eligible for assistance; do the financial resources provided allow the recruiting and retention of staff with the desired qualifications?

Factors to Consider in Selecting a QRIS Evaluation Design

This section reviews four key factors that should be considered in choosing an evaluation design: (1) strength of evidence required to address research questions and program improvement inputs needed to inform program management, (2) stage of QRIS development, (3) available funding, and (4) the timeframe in which research questions must be answered. It is important to keep in mind that certain types of evaluation designs are only possible with certain QRIS configurations; design of the initiative and of the evaluation are best conducted in concert. In addition to the evaluation design, another consideration is who will conduct the evaluation. Table 1 summarizes key factors to consider when planning an evaluation and describes the relative advantages and challenges of each.

Level of Evidence and Usefulness for Program Improvement

QRIS evaluations may focus on *whether* a policy improvement initiative meets its objectives (referred to as an impact or outcome evaluation and characterized by "A" questions above) or *how* an initiative works (referred to as a process or implementation evaluation and characterized by "B" questions). Policy makers and funders are usually interested in funding impact or outcome evaluations because they want evidence that public and private investments are paying off. Process/implementation evaluations are designed to inform and improve program development and operations. These research designs are not mutually exclusive; indeed, many rigorous and comprehensive evaluations combine an impact/outcome study with a process/implementation study. This combination can be especially powerful during the piloting and scale-up stage because it provides a rich picture of what is working and helps to explain why certain outcomes were observed for certain groups.

Selection Bias occurs when the participants (provider organizations, staff members, communities) self-select or are selected based on criteria which may be related to the probability of attaining a certain outcome. For example, the design of a QRIS evaluation may result in the most motivated provider organizations, staff, or parents participating in the QRIS. Under these circumstances, it is impossible to determine whether outcomes are due to the supports and incentives offered, or to the higher motivation and engagement of the recruited participants. Such evaluation designs cannot determine whether system-wide improvement is occurring. The objective of improving the lowest quality providers may also be untested. Random assignment is the classic antidote.

Impact evaluations should rely on rigorous evaluation designs that enable causal statements such as, "The QRIS significantly improved children's school readiness." Outcome evaluations that depend on less rigorous evaluation designs enable statements such as, "The QRIS is associated with improvements in children's school readiness;" a causal association cannot be made. Although differences in these two types of designs may appear subtle, the different level of scientific rigor each provides affects the credibility of the findings and the conclusions that can be drawn. For example, some less rigorous evaluations have documented a rise over time in the numbers of providers achieving higher QRIS ratings as well as increases in observed quality and improved child outcomes. However, the research designs on which the evaluations are based do not provide evidence to support the inference that participation in QRIS *caused* these associations; other unmeasured factors may have been driving them.

Experimental and quasi-experimental evaluations provide the highest level of evidence and allow causal statements about QRIS effects. Experimental and quasi-experimental evaluations are most likely to detect effects such as improvements in targeted short-term outcomes (e.g., workforce skills and attitudes, parent use of information to guide care selection, the quality of care) as well as long-term outcomes such as school readiness. To find valid effects, it is important to be able to compare the QRIS group to something else; these comparisons depend on an assumption of initial equality between the groups. An experimental study ensures equality by randomly assigning some child care providers to a treatment group that receives the QRIS intervention and others to a comparison group that does not. Following successful random assignment, the two groups may be assumed to be similar at the start of the evaluation; any differences observed at the end of the evaluation can be presumed to be *caused* by the QRIS initiative. Quasi-experimental designs *match* providers or communities on important characteristics at the start of the evaluation (such as the number of children served and proportion of children receiving subsidy) and then assess changes in outcomes for those that receive the QRIS intervention and those that do not. A number of current evaluations follow a cohort of children enrolled in settings participating in QRIS and compare their progress to a group of children who were not enrolled in such settings. As noted above, if these studies do not match providers or communities on key characteristics, causal statements are not possible using such designs, and they are less likely to yield valid estimates of differences across groups. Nevertheless, these studies can meet stakeholder needs for information about how children are faring. Evaluation designs that include baseline data collected prior to random assignment and the start of the QRIS intervention and follow-up data collected after the intervention has had time to mature can account for any baseline differences between the groups and increase the power and precision of the impact analyses. Tradeoffs between evaluation rigor and costs are described below.

Implementation/process evaluations document how a QRIS operates and inform program improvement. Although not designed to answer questions about whether a QRIS "works" in terms of desired outcomes, well-planned and rigorously conducted implementation research provides insights into whether all components of a QRIS are operating as intended and identifies areas that need improvement. Depending on the QRIS logic model and the guiding research questions, implementation evaluations may document how closely the implemented policy adheres to the QRIS design in terms of dosage, content, and uptake of technical assistance and supports, the proportion of child care providers and children in QRIS-rated settings, and staff turnover at all levels. The results of implementation evaluations are often critical in identifying problems that need correction before a fair test of impacts is possible. These findings may also suggest why better results are found for some communities or providers than others.

Simply selecting a rigorous design is not enough; stakeholders, administrators and evaluators must uphold the evaluation's requirements. There are many different ways to compromise even the most rigorous evaluation design and undermine the quality of an evaluation. For example, providers, advocates or evaluators may not be comfortable with conducting a random assignment process that leaves many low-quality providers unassisted and decide to move some low-quality providers into the treatment group. Allowing such "crossovers" undermines random assignment and the ability of evaluators to draw valid conclusions from the study findings. Administrators may reduce the planned supports and incentives in the face of budgetary constraints. Center or program directors may be reluctant to share detailed quality ratings or evaluation findings with their staff and with parents of enrolled children. A combination of stakeholder participation in the QRIS design phase and ongoing outreach and education is necessary to maintain engagement, cooperation, and fidelity of the evaluation design.

Stage of QRIS Development

Evaluation need not be deferred while a program is being established. Indeed, programs can benefit from QRIS evaluation findings at all stages of QRIS development. But it is important that the evaluation design align with the QRIS's stage of implementation. This section presents examples of evaluation research questions and design options tailored to three different QRIS development stages: (1) pilot and scale- up, (2) early operation (the first two to five years), and (3) mature operation (more than five years).

The QRIS pilot and scale-up stage allows for testing of the QRIS approach as well as assessment of changes in targeted outcomes. Many states and municipalities conduct evaluations of their initial QRIS implementation and efforts to move from a relatively small-scale pilot to statewide implementation. Important implementation research questions should be guided by the system's underlying logic model. Using Zellman et al.'s and Brandon's logic models, presented above, they may include: are more parents learning about the system? Do incentives cover the cost of quality improvement? What do key stakeholders view as the major implementation challenges and successes? An implementation study design that includes triangulation of information from multiple sources (program developers, operators, staff, and parents; data systems that track participation and service receipt; and community surveys about QRIS knowledge) provides data crucial to improving services and community awareness of QRIS.

Experimental evaluation designs used during a pilot or scale-up stage often capitalize on scarce system resources, i.e., not all providers who volunteer can be accommodated. Random assignment of willing providers to (1) a group that receives the QRIS ratings, technical assistance, and other quality improvement supports, or to (2) a comparison group that does not receive the supports, allows for fairness in selecting participants and a strong research design. Such a pilot can assess the impact of QRIS on short-term outcomes (e.g., after six to nine months) and identify needed adjustments to the intensity of the improvement strategy, for example, the number of technical assistance hours per provider or the level of financial incentives at each rating level. If an experimental evaluation finds short-term positive impacts on quality, other long-term outcomes may be expected. If quality is not affected by the QRIS intervention, it is important to understand why. If resources are inadequate or misdirected, corrections may help to boost quality.

A challenge in using random assignment of willing providers is that while it can measure impacts on individual providers, it cannot measure impacts at the market or system level. An alternative is to randomly select among counties or other jurisdictions and offer supports and incentives to all providers and parents within those jurisdictions. The degree to which the supports and incentives offered by a QRIS can attract voluntary provider and parent participation is often an important question for evaluation.

One issue related to conducting an impact evaluation during the pilot or scale-up phase is that some may view it as too early to draw conclusions about whether the QRIS is working. The argument is that pilots take time to reach a level of intensity and quality that could affect outcomes. Although this may be true, the potential lessons learned about how implementation may be improved may outweigh these concerns. One approach to resolving disagreements about the scope of an early evaluation is to seek consensus among stakeholders that the intent of an evaluation at this stage is to inform QRIS improvement rather than make up or down decisions about whether it should continue.

The early operation stage (first two to five years) provides the opportunity for large-scale, longitudinal evaluation. Research questions at the early operation stage may focus on assessment; data may be collected to answer questions about child outcomes as well as aspects of the QRIS required to maintain the intervention. Outcome questions aligned to the logic model might include: How has overall quality changed over time? Are children in the locality better prepared for kindergarten and school success? Outcomes studies may track changes in observed quality and analyze trends prior to the start of QRIS and in the years following establishment of the system.

Implementation research questions at this stage may include: How do ECE –SAC providers perceive the incentives—do they believe the effort required to reach the next rating level worth the incentives offered? Are parents using the ratings to make ECE-SAC choices? Are children in greatest need of improved quality care enrolled in QRIS settings? During this period, program stakeholders and operators may also seek information about the cost of QRIS and look to evaluators for data about the financial and in-kind costs of operating the system.

Generally, evaluations conducted at this stage that include in-person quality observations, direct child assessments, parent surveys, and qualitative interviewing and focus groups require sampling of participating child care businesses, families, and the children in care because there are too many to include all of them. Evaluators must specify their approach to selecting study participants to ensure that the findings represent the population. A design that includes interviewing and assessing only those that are easiest to reach ("a sample of convenience") are not rigorous, are subject to selection bias and manipulation, and should not be used.

The mature operation stage (five years and beyond) provides the opportunity to assess trends in quality improvement and outcomes and use data for program improvement. As described in the previous section, documenting trends in ECE-SAC quality and targeted outcomes is one approach to evaluating existing, full-scale QRIS implementation. An interrupted time series design involves comparing several years of data on quality or child and family outcomes from before the implementation of full-scale QRIS to the same outcomes assessed over several years after implementation (the more time periods that can be included, the better). This approach documents trends over time but does not rule out the possibility that factors external to the QRIS (for example, large increases or decreases in state and federal child care investments) explain or contribute to these trends. Nevertheless, a number of mature QRISs use this type of analysis in response to funder reporting requirements. Trend analyses require good data collected at a reasonable point in time, e.g., kindergarten readiness assessments or early elementary standardized test scores. These types of analyses require access to existing state datasets and rely on the quality of those data.

Even in mature systems, analyses can inform deployment of additional training and technical assistance resources. For example, with QRIS data collected over time about specific providers, QRIS quality improvement staff can help target additional technical assistance and training to providers not showing quality increases over time. For example, if a group of 2-star centers in a given geographic area seems stuck at that level, follow-up may help to identify reasons for stagnation, e.g., incentives are inadequate to cover quality improvements; providers are able to fill all available spaces with a 2-star rating.

A mature QRIS may also benefit from reassessment of its quality standards and approach to assessing quality to ensure that they reflect recent research findings and that they foster quality changes most linked with targeted child outcomes. Although providers often resist changes to the standards, sticking with an outmoded rating system may limit the effectiveness of a QRIS in enhancing quality. Refreshing the standards and reevaluating how the rating areas are weighted ensures that the best of the leading research evidence is incorporated into mature QRISs. Initial standards often represent a compromise between the level of quality desired and what seems feasible given a poorly educated and compensated ECE-SAC workforce. As quality improvements occur, shifting the entire rating scale upward may become both desirable and feasible. Such changes, however, may pose some evaluation challenges, as comparability over time is lost.

Partnership between administrators and evaluators at this stage may also help identify important system-level issues that should be addressed, for example trends in the quality of care provided to children receiving subsidized care. By linking state QRIS service use data and subsidy receipt data at the level of the individual child, policy makers and program operators can explore whether key QRIS goals, e.g., improving the quality of care for the most at-risk children, are being met. Many states develop QRIS specifically to encourage parents of children at risk for school failure to choose child care settings of better quality. If a state provides tiered reimbursement for children on subsidy, providers may seek to enroll these children at higher rates than they would have without this QRIS incentive. Linking data on child-level subsidy receipt to the quality of children's settings provides information program operators can use to improve their QRIS.

Table 1: Summary of Evaluation Design Decisions—Inherent Advantages and Challenges to Consider

Design Consideration	Advantages	Challenges to Consider
Type of Evaluation		
Outcome evaluation	Best way to assess QRIS impact on ECE-SAC quality and child outcomes	May yield "false negatives" early on; may be hard to explain modest outcomes or how to improve
Implementation/Process evaluation	Best way to collect information about provider and parent experiences, and inform program improvement	Does not provide rigorous evidence about outcomes; potential for "response bias" as participants put best light on experience or voice criticism based on unrealistic expectations
Outcome Evaluation Design		
Experimental/RCT	Most robust evidence of change; most likely to yield valid estimates of effects	Requires withholding support to some families, providers, communities
Quasi-experimental	Can capitalize on existing variation in availability of services.	Harder to distinguish impact of QRIS from other factors
Unit of Analysis		
Individual-child staff member	Greatest ability to observe variations in impact on different child and provider subgroups	Most costly data collection and verification
Provider, community	Less data collection; easier to use administrative data for evaluation; focuses attention on provider quality	Harder to distinguish variations in impact
System or market-level	Captures broader effects, e.g., changes in supports and incentives, changes in overall level of quality	Evaluation compares counties or communities; difficult and costly to offer supports and incentives to all providers and families.

Design Consideration	Advantages	Challenges to Consider
Evaluation Timing		
Developmental period	Capture lessons early; help system improve	Lack of outcomes may be misinterpreted as failure
Mature period only	Ability to demonstrate outcomes over time, different ages of children	Without early study, may miss system improvement opportunities; unable to interpret any differences from prior period(s)
Type of Evaluator		
Internal	Best access to existing administrative and outcome data; can assess need for new data; can integrate results into policy decisions in real time	May lack specialized skills; agency staff may be diverted to other tasks or feel pressure to minimize negative results; not independent, objectivity can be questioned
External	Higher credibility with skeptics; brings specialized skills and experience in evaluating other QRISs	May be more expensive; does not build public agency capacity, so agency may not be able to carry on evaluation after initial contract period

Available Funding

Funding limits the ability to conduct rigorous, longitudinal evaluations of QRISs. Often states underestimate the cost of evaluation and inadvertently jeopardize the quality of the research that can be done by failing to set aside sufficient evaluation resources. It is important for states to consider the costs of evaluation as it seeks QRIS funding. Setting aside sufficient evaluation funds will enable a more rigorous evaluation, which will increase the odds that states will learn which activities are most effective in achieving QRIS benchmarks and goals.

Alignment of the design options with the available funding helps both the state and prospective evaluators set expectations for the work. The relative costs of different evaluation approaches may vary considerably and depend to at least some degree on the level and type of data already being collected as part of the QRIS (e.g., if observations are made of many classrooms on a regular basis, the evaluation design can capitalize on these data so data collection costs are reduced).

States and federal government grants and contracts are not the only sources of QRIS evaluation funding. Many states have established public-private collaborations that support QRISs; funding for QRIS evaluations also has come from foundations. Creative approaches to combining funds from different sources may allow states to afford evaluation designs that are otherwise out of reach. Many evaluators will work with states to secure the funding for an evaluation.

Evaluation Timeframe

Policy makers and other state stakeholders seek information and data to guide decision making in the near term while many of the most rigorous evaluation designs collect longitudinal data about QRIS impacts and take a number of years to complete. As described above, there are creative alternatives to conducting large-scale, longitudinal studies that require intensive data collection. These include using existing or expanded state data. QRIS developers, managers, and other stakeholders must work together to educate state leaders and advocates about why investments in QRIS evaluation and data to guide program management are important and worth the wait. Evaluators can help states consider their information needs and design multi-method studies that provide both short-term implementation and output data and long-term outcome data to address accountability and program improvement needs. Thinking through the options and timing of data needs is an important step in planning an evaluation and can be accomplished through consultation with other states conducting research or as part of a research design contract.

Who Should Conduct an Evaluation?

As state administrators consider the factors described above, it is important to weigh the benefits and challenges to using different types of evaluators. The evaluation could be developed and conducted "in-house;" there could be "in-house" implementation of a design developed through collaboration with an outside evaluator who might also help to obtain federal or private funds for the evaluation (more of a hybrid approach), or there could be a contract for an outside evaluation. Before deciding which approach to pursue, issuing a request for proposals (RFP) or designing an internal evaluation, it is critical to review or develop a logic model and clearly lay out the research questions inherent in that model that the state wants to answer and to identify the budget available to conduct the evaluation. There are many ways to align state questions, resources and evaluation design options, from consulting with other states to funding a preliminary design contract with seasoned QRIS researchers. A design contract may be the most efficient way to bring in QRIS evaluation experts who can work with stakeholders to identify research questions, document the QRIS logic model, and cost out different evaluation designs. At this stage, they can also help the state develop the terms and level of detail to be included in an evaluation RFP. A design contract or consulting agreement may be arranged for a moderate cost depending on who conducts the research and the range of tasks required. States should seek experts who are able to offer a range of expertise that includes developing sampling plans for experimental studies to planning key informant and focus group discussions with staff and parents. Usually a design contract can be completed in six months; the time to conduct the evaluation will vary depending on the nature of the research design.

Internal or External Evaluator?

Is it better to hire an independent evaluator, or to conduct an internal evaluation? Each has advantages. An internal evaluation generally can more easily access data and integrate results into policy decisions; an external evaluation brings higher credibility, greater expertise, and experience. An outside evaluator needs to be paid; internal evaluation costs are real but less obvious. The decision may depend on available capacity, budget, and credibility needs.

Conclusion

A good evaluation design, thoughtfully developed, can provide information critical to improving the system at many points in the process, and increase the odds of its ultimate success. Evaluation is unquestionably challenging, but no more so than the launch and operation of a QRIS. The networks and references in next section can help states develop a deeper understanding of evaluation approaches and plan and execute QRIS evaluations that address stakeholder and system needs and produce timely and valuable information.

Resources and References

Resources

Child Care & Early Education Research Connections (2008). **Quality Rating Systems: A Key Topic Resource List.** New York: Child Care & Early Education Research Connections. http://www.researchconnections.org/files/childcare/keytopics/QualityRatingSystems.pdf

An annotated bibliography of selected research focused on the design, implementation, and evaluation of Quality Rating Systems and Quality Rating and Improvement Systems in early childhood and after school settings.

QRIS National Learning Network

http://qrisnetwork.org/

The Network provides information, learning opportunities, and direct technical assistance to states that have a QRIS or that are interested in developing one. Its *National Resource Library* assists states in learning more about QRIS and their elements and in QRIS planning. The library contains toolkits, handouts and published documents on a variety of searchable topic areas.

The Network's *State Resource Library* contains detailed QRIS implementation information, including training guides, forms, and technical assistance materials that individual states have developed for their QRIS.

State QRIS Contacts who have agreed to serve as peer resources for one another are listed, as are *Technical Assistance* providers.

Quality Rating & Improvement System Resource Guide

http://nccic.acf.hhs.gov/qrisresourceguide/

Developed by the National Child Care Information and Technical Assistance Center, the Resource Guide is a Web-based tool for states and communities to explore key issues and decision points during the planning and implementation of QRIS. The guide is divided into eight sections, which cover topics ranging from the initial design process to evaluation. In addition, each section includes a set of questions for users to consider and discuss when planning, implementing, or revising QRIS. The guide provides state examples throughout to illustrate strategies used to develop and implement QRIS, as well as selected resources. It also includes an interactive map that links users to information about the status of each state's QRIS or other large-scale quality improvement initiative.

The Child Care Quality Rating System (QRS) Assessment

Tout , K., Starr, R., Soli, M., Moodie, S., Kirby, G. & Boller, K. (2010).**The Child Care Quality Rating System (QRS) Assessment: Compendium of Quality Rating Systems and Evaluations, OPRE Report.** Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

http://www.childcareresearch.org/childcare/resources/18554

Describing 26 Quality Rating Systems nationwide (19 statewide and 7 local or pilot), the Compendium presents comprehensive information through cross-QRS matrices and individual QRS profiles. Eighteen of the 26 had undertaken some type of evaluation by the time of data collection. Matrices compare their approaches to evaluation; the main categories of research questions were: system implementation (asked by 7), validation of quality ratings (7), improvements in program quality (9), and child outcomes (4). Individual QRS profiles list their published evaluation reports.

Lugo-Gil, J., Sattar, S., Ross, C., Kirby, G., Boller, K. & Tout, K. (forthcoming). **The Quality Rating and Improvement System (QRIS) Evaluation Toolkit, OPRE Report.** Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

The QRS Assessment Toolkit will provide guidance, recommendations and evaluation support on a range of topics including: development of a logic model and research questions, evaluation design and methods, and selection of measures.

References

[1]Boller, K., Del Grosso, P., Blair, R., Jolly, Y., Fortson, K. N., Paulsell, D., ... Kovac, M. C. (2010). *The Seeds of Success modified field test: Findings from the impact and implementation studies*. Princeton, NJ: Mathematica Policy Research.

[2]Brandon, R. (forthcoming). *Financing to Promote Quality in Early Care and Education and School-Age Care: Incentives, Supports and Affordability,* Research-to-Policy, Research-to-Practice Brief. Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

[3]Breitner, L., Brandon, R. & Lalic, N. (2010). *Budgeting as a tool for policy development. Training materials prepared for the UNICEF Social Protection and Inclusion Project in Bosnia and Herzegovina.* Retrieved from: http://www.hspc.org/topics/SPIS/Module1_Background_Materials.pdf

[4]Chen, H.T. (2005). *Practical program evaluation: Assessing and improving planning, implementation, and effectiveness*. Thousand Oaks, CA: Sage Publications.

[5]Elicker, J. & Thornburg, K. (2011). *Evaluation of Quality Rating and Improvement Systems for Early Childhood Programs and School-Age Care: Measuring Children's Development*, Research-to-Policy, Research-to-Practice Brief OPRE 2011-11c. Washington, D.C.: Office of Planning, Research and Evaluation, Administration for Children and Families, U. S. Department of Health and Human Services.

[6]Hatry, H. P. (2006). *Performance measurement: Getting results, 2nd edition*. Washington, D.C.: Urban Institute Press.

[7]Rossi, P. H., M. W. Lipsey & H. E. Freeman (2004). *Evaluation: A systematic approach*. Thousand Oaks, CA: Sage Publications.

[8]W.K Kellogg Foundation. (2004). *Logic model development guide*. Battle Creek, MI: W. K. Kellogg Foundation.

[9]Zellman, G. L., M. Perlman, V.-N. Le & C. M. Setodji (2008). *Assessing the validity of the Qualistar Early Learning Quality Rating and Improvement System as a tool for improving child-care quality*. Santa Monica, CA: Rand Corporation (MG-650-QEL).