State-Funded PreK Policies on External Classroom Observations:

Issues and Status

Policy Information Report

ETS

Listening. Learning. Leading.®

This report was written by:

Debra J. Ackerman
Educational Testing Service

Author contact:
dackerman@ets.org

Additional copies of this report can be ordered for $15 (prepaid) from:

Policy Information Center
Mail Stop 19-R
Educational Testing Service
Rosedale Road
Princeton, NJ 08541-0001
(609) 734-5212
pic@ets.org

Copies can be downloaded from:
www.ets.org/research/pic

February 2014
Policy Evaluation and Research Center
Policy Information Center
Educational Testing Service

Table of Contents

Preface

Early education programs are increasingly being promoted by states and the federal government as an integral part of their efforts to ensure that all children enter school ready to learn. As these programs and their enrollments have grown in recent years, so too have efforts to monitor their quality and performance. A common focus is on documenting the quality of children's learning experiences through the collection of classroom observation data. In order for these data to be useful for informing the monitoring process, however, they need to demonstrate evidence of being appropriate and defensible for their intended interpretation and subsequent uses.

In this new Policy Information Report, Debra Ackerman examines the variety of state PreK classroom observation policies on program decisions that are informed by observation score data, the protocols being used, and how often such data are collected from classrooms. Just as important, the author reminds us of the particular validity and reliability challenges that are inherent in relying on classroom observation score data for a variety of low- and high-stakes decisions.

It is our hope that this report will cause policymakers, school leaders, and practitioners to reflect on their early education program classroom evaluation policies, whether they meet acceptable levels of validity and reliability, and what actions they can take to improve the usefulness of data collected to improve the quality of children's early learning experiences. As federal and state efforts to improve access to high quality early education continue to grow, it will be increasingly important to monitor this critical segment of our education pipeline.

Michael T. Nettles
Senior Vice President
Policy Evaluation and Research Center, ETS

Acknowledgments

The author wishes to acknowledge the help of many people in producing this report, including the state PreK representatives who graciously participated in the study's survey. In addition, the report was reviewed by Andrea DeBruin-Parecki, Drew H. Gitomer, Daniel F. McCaffrey, Robert C. Pianta, Donald Powers, Richard Tannenbaum and Cindy Tocci. Eileen Kerrigan was the editor. Marita Gray designed the cover and Sally Acquaviva provided desktop publishing. Errors of fact or interpretation are those of the author.

Introduction

Classroom observation data can serve as an important component of early education quality improvement and accountability efforts.[1] Perhaps not surprisingly, such data increasingly are being used to inform key decisions in programs serving young children. For example, the state-level departments that regulate and administer their respective publicly funded Pre-Kindergarten (PreK) programs use these data as part of the ongoing monitoring process.[2] Formal observations of federally funded Head Start classrooms help determine which grantees may be subject to open competition and thus potentially lose their funding.[3] The recent Race to the Top – Early Learning Challenge (RTTT-ELC) competition gave priority to applicants that proposed strengthening their system for rating and monitoring the quality of early education programs.[4] Classroom observation scores also contribute to the tiered quality rankings of — and subsequent level of child care subsidy reimbursements provided to — child care settings that participate in state Quality Rating and Improvement Systems (QRIS).[5] And, a suite of observations inform the quality improvement process in the public and private partnership-funded Educare schools for children ages 0–5 throughout the United States.[6]

While observation score data already contribute to a variety of low- and high-stakes early education decisions, because of the larger K–12 focus on the measurement of teaching quality,[7] one might argue it only is a matter of time before these data have an even greater number of consequential implications. Moreover, K–12 teaching effectiveness decisions can drive both staff salary increases/bonuses and pay freezes or job termination.[8] It therefore is particularly important that policymakers and stakeholders know they can "bank on" the quality of observation score data when making important decisions about a teacher, program grantee, school, or overall early education program.

Three broad factors contribute to the extent to which the use of classroom observation score data is valid for informing any decision: the observation protocol(s) used, the capacity of observers to generate reliable score data, and how frequently data are collected from any one classroom. Because each of these factors presumably is governed by related policies, such policies are equally important contributors. However, just as the judgments that rely in part on classroom observation data may have consequential outcomes for early education staff and the settings in which they work, the policies governing these three factors will have implications for early education program resources. As a result, policymakers likely will need to consider how their optimal classroom observation score data requirements might be balanced against the realities of their program's context and capacity.[9]

Given the variety of publicly funded early education programs relying on observation score data, as well as the continued pressure on these programs to contain or cut operating budgets,[10] this larger context is particularly salient as the field expands its discussion regarding best classroom observation practices. To help highlight some of the prospective agenda items for that discussion, this report summarizes the emerging — and often less-than-definitive — literature base on the potential validity and reliability issues related to policies on classroom observation protocols, observer capacity, and frequency of observation data collection. As the reader will notice, while some of this literature arises from early education-focused research, the majority of existing research on these topics is situated in middle and high school classrooms. In addition, in light of President Obama's proposal to expand 4-year-olds' access to high-quality, publicly funded preschool,[11] the report also examines the variety of state-funded PreK program policies on classroom observations as a means for illustrating why these validity and reliability issues should be considered as part of the best practices discussion agenda.[12]

To set the stage for this dual inquiry, the report begins with an overview of state-funded PreK in the United States. This is followed by a brief description of the typical classroom observation process and two widely-used, preschool-focused classroom quality

measures. After describing the current research base on the reliability and validity issues related to choice of observation protocol, ensuring observers have the capacity to generate reliable scores, and determining how frequently score data will be collected, the report turns to the status of PreK observation policies. The report concludes with some general perspectives on potential agenda items for future best practices discussions.

In his 2013 State of the Union address, President Obama proposed partnering with states over a 10-year period to expand 4-year-olds' access to high-quality, publicly funded preschool.[13] If funded, this effort will build on enrollment increases in current state-funded Pre-Kindergarten (PreK) programs that have occurred over the past decade. For example, during the 2001–2002 school year, 581,705 4-year-olds, or 14.8 percent of the entire population in this age group, were enrolled in 45 PreK programs across 40 states.[14] A decade later, the number of 4-year-olds participating in these programs jumped to 1,151,653, representing 28 percent of all 4-year-olds. The number of PreK programs across the United States increased during this period as well, with over 50 different state-funded initiatives in 40 states and the District of Columbia in operation in 2011–2012. This growth is particularly impressive given recent state education budget constraints.[15]

Many factors have contributed to the growth of interest and enrollment in publicly funded PreK, including a widening focus on improving preschoolers' kindergarten readiness and, in turn, their long-term academic outcomes and ability to contribute to the nation's economic growth.[16] Another important contributor has been an increase in the participation rates in such programs of school districts, child care centers, and Head Start grantees (referred to in this report as *individual providers*). As this mix of individual providers has expanded, to help ensure that the quality of PreK enrollees' experiences does not vary based on classroom setting, state policymakers have established program and learning standards and monitoring policies.[17] The collection of monitoring data can provide state and local PreK administrators with the capacity to engage in a cycle of planning, doing, reviewing, and acting,[18] as well as support reflection on what standards and/or programmatic inputs need to be revised as a means for meeting a PreK initiative's goals.[19] And, depending on the research design, data from monitoring efforts also has the potential to support program accountability or quality improvement efforts.[20]

While most PreK programs have monitoring policies, how monitoring data are used varies. In some cases, these data contribute to "low-stakes" program improvement purposes, including determining the professional development teachers need. Other PreK programs use monitoring data to inform high-stakes decisions, such as whether individual providers are in need of corrective actions or sanctions. Some PreK programs rely on monitoring data to implement changes in preschool policies.[21]

As might be expected given these diverse purposes, PreK programs report the collection of a variety of monitoring data, including documentation of children's learning outcomes.[22] Another source of data is PreK classroom observation scores.[23] However, as is the case with child assessment results, if these data are to inform the monitoring process, they need to demonstrate evidence of being appropriate for their intended use(s) and subsequent interpretation. One approach to gathering such evidence is through an examination of the protocol used, the capacity of data collectors to reliably use the protocol, and when and how frequently data are collected. To provide some context for the importance of these topics, the general process of collecting observation data in early education classrooms is described next.

To help explain the reliability and validity challenges that come with the reliance on observation score data, it is useful to understand the typical procedures for administering such measures in early education settings. For example, while each measure presumably will have its own specific instructions, the process typically involves an observer sitting, standing, and/or walking around a classroom while looking for and noting evidence of a specific aspect of teachers' practice or children's classroom experiences. The observer also may ask to see a teacher's lesson plan or interview her about typical teaching practices. The exact practices and/or experiences being noted are guided by the items in the observation protocol. Then, as the observer collects evidence for an individual item or larger domain or subscale of the instrument, he or she compares it to the information in the scoring rubric, as well as any other developer-provided notes or tools, to determine the appropriate score. The scores then are tallied or averaged in some way to determine an overall score.

As an example of the practices or experiences that are targeted as part of an observation, one item of interest might assess the degree to which sand and water play are available in a classroom serving 4-year-olds. The observer then would focus on all of the relevant components in the sand/water table area, as opposed to counting and categorizing the puzzles or writing-related materials that are accessible to children. To determine the appropriate score for this item, he or she might need to consider such variables as the size of the table(s), how many children can be accommodated at any one time, the quantity and variety of sand or water toys, and whether the depth of the sand or water is adequate for using the toys. Another criterion might be the amount of time children use the sand/water table area each day.

Another related, yet more complex item might focus on the degree to which a teacher uses discussions, experiments, and higher-order thinking to help her 4-year-old students develop their early science and mathematics skills while using the sand and water tables. In this case, instead of merely tallying quantities, such as the number of sand and water toys or the depth of the sand and water, the observer would key in on the activities students are asked to undertake while at the sand or water tables. Also of interest would be the interactions the teacher has with children while engaged in these activities that also support their learning in these content areas.

Two well-known measures present examples of the different types of evidence gathered and scored as part of the early education classroom observation process. First, the *Early Childhood Environment Rating Scales – Revised (ECERS-R)*[24] provides a global overview of programmatic features found in center-based settings serving children ages 2 ½ to 5 years old. Its 43 items are categorized within the seven subscales of Space and Furnishings, Personal Care Routines, Language-Reasoning, Activities, Interactions, Program Structure, and Parents and Staff. The original ECERS[25] and the more recent ECERS-R have a long history within the early care and education field due to the protocol's role in numerous research projects and state QRIS efforts.[26]

A second well-known observation protocol is the *Classroom Assessment Scoring System (CLASS)*: *Pre-K*,[27] which focuses on teacher-child interactional processes in classrooms for children who are 3 to 5 years old. Its 10 items are referred to as "dimensions" and are organized under the three domains of Emotional Support, Instructional Support, and Classroom Organization. The CLASS Pre-K is used in Head Start settings to determine which grantees should be subject to open competition.[28]

While their focus is different, both protocols are similar in that they use a Likert-like scoring scale of 1 to 7. The developers of each protocol further categorize their respective numerical scales into low-, mid-, and high-range quality scores. For example, ECERS-R scores of 1–3 indicate inadequate to minimal quality, 4–5 indicate good quality, and 6–7 indicate excellent quality.

On the CLASS Pre-K scale, scores of 1–2 are considered to represent low-range quality of teacher–child interactions, 3–5 are mid-range, and 6–7 are considered high-range quality.

When considering the two sand and water table examples, if included in an observation protocol and the classroom did not contain a sand or water table, the first example's related item presumably would be scored in the inadequate or low-quality range. Conversely, if the classroom contained large sand and water tables that contained numerous toys which are available to children on a daily basis, the item most likely would receive a score that is indicative of one of the higher quality categories. For the second example, assuming that the sand and water tables were available for use in the classroom, a higher quality range score might reflect the teacher asking students to test out hypotheses regarding the amount of sand or water that are needed to balance out a scale, or how many smaller cups of sand will equal the amount held by a larger container. Also potentially contributing to this score would be the complexity of the discussions that the teacher has with children as they undertake these hypothesis-testing activities.

Regardless of what range of scores are generated by classroom observations, if early education policymakers aim to generate scores that can reliably inform a "plan, do, review, and act" model, they first must determine a set of key protocol, observer, and data collection frequency policies. These policies, as well as their related validity and reliability challenges, are discussed next.

Classroom observation score data have the potential to inform policymakers' and other stakeholders' decisions for programs serving young children. However, to generate the type of data that can be most useful for the decision-making process, policymakers also need to be aware of key validity and reliability issues when crafting their classroom observation policies. *Validity* refers to the extent to which the accumulation of evidence collected as part of any assessment supports the interpretation of its scores for a particular purpose and for a specific population. This might include informing the type of professional development teachers need or making a judgment about an individual provider or teacher for accountability purposes. *Reliability* refers to the degree to which any variation in scores reflects differences in classroom quality as measured by a particular protocol's rubric, rather than the accuracy of the observers' judgments in using the protocol.[29] In short, if two or more raters observe the same classroom at the same time, their ratings or scores should not vary significantly.

In an ideal world, an early education program's policies will support the large-scale collection of classroom observation data that will effectively inform myriad decisions regarding individual teachers and providers, as well as the larger early education program in which they participate. However, no one-size-fits-all, gold-standard method for accomplishing this complex goal exists. Furthermore, any large-scale data collection effort typically involves context-specific tradeoffs between what may be ideal and what is feasible given time, budget, and/or personnel resources.[30]

Yet, this does not mean that any set of policies is sufficient for supporting the collection of data that will accurately inform both low- and high-stakes decisions. Instead, programs need to seek out the policy "sweet spot" that balances their information needs, context, and resources.[31] When the goal of a large-scale data collection effort is to generate observer-generated information on classroom quality, three key issues need to be considered: the observation protocol(s) to be used, the capacity of observers to generate reliable score data, and the frequency with which observation data will be collected from any classroom.

Choosing an Appropriate Observation Protocol

The first key issue to consider when crafting early education classroom observation policies is: Which protocol(s) should be used? Responding to this question is no small task, as recent reviews of all of the available observation-based measures designed for classrooms serving young children show that there are at least 50 protocols from which policymakers may choose. Moreover, the focus of individual protocols can be quite different, with some measures assessing easily quantifiable inputs such as teacher-child ratios, classroom square footage, and access to child-sized sinks and toilets, but others examining more difficult-to-quantify teacher-student interactions or the types of activities available in the classroom. Additional protocols estimate how much time children spend engaged in certain activities or the extent to which a specific curriculum is implemented.[32]

Alignment with purpose and setting. Recalling that validity refers to the extent to which the interpretation of any assessment's scores is appropriate for a particular purpose, a prime consideration when choosing an observation protocol is its alignment with the task at hand.[33] For example, a PreK's monitoring goal may be to determine if its teachers need professional development related to enhancing children's learning and skills in math and science. If so, the observation protocol should focus on the different interactions teachers have with children to support their mathematical and scientific thinking (e.g., "What will happen if we add one cup of sand to the blue side of the scale, but leave these three sand toys on the red side of the scale?"), as well as the classroom materials and activities available that contribute to children's math and science learning (e.g., books that talk about differences in objects' weight, scales, and measuring cups).[34] However, another goal may be to evaluate the overall quality of a specific classroom or individual

provider. In this case, a protocol focused solely on math or science most likely will provide too narrow of a picture, particularly if one of the goals of the larger PreK program is to enhance children's outcomes in a variety of academic and developmental areas.

Because validity also is related to the relevancy of score data for a specific population, next to be considered is the early education setting in which the observations will take place, including the ages or characteristics of the children enrolled in those settings. The reason this aspect of validity also is critical is that although the majority of these 50-plus protocols are aimed at child care centers or K–12 school-based preschool classrooms, others specifically focus on the quality of home-based family day care settings.[35] In addition, some protocols target teaching practices in classrooms serving children 0–5 years old or 3–8 years old, while others are designed to measure practices that pertain only to infants and toddlers. Also to be considered is whether an observation measure is designed to capture those aspects of classroom quality that support the learning and development needs of children who are dual-language learners or have disabilities.[36]

Because no single classroom observation protocol can adequately cover every early education setting and/or age group, much less all facets of early education classroom quality, an additional issue to consider is how many different observation protocols should be used to inform a monitoring processes' goal(s). On the one hand, using just one observation measure can provide decision makers with uniform data regarding teachers, individual providers, and/or the overall early education program.[37] However, if a program has multiple monitoring goals, it is unlikely that one measure will be adequate.[38] In addition, if classrooms vary in terms of teacher qualifications, the demographics of the children served, or curriculum used, relying on score data from just one protocol can produce data that are less useful for a variety of decisions.[39] Also, because individual protocols typically evaluate very different aspects of early education quality, classrooms or teachers can be categorized as good, high quality, or effective when using any one measure, yet also be low quality or ineffective as defined by an additional measure.[40]

Psychometric properties. Once policymakers narrow down the list of potentially appropriate observation protocol choices, it is important to note that the mere claim of alignment with a purpose and/or setting and population does not necessarily mean the protocol actually measures the constructs of interest in a reliable manner, much less should be used to draw a particular conclusion.[41] Furthermore, policymakers also must investigate the extent to which a measure's score data consistently will support the judgments to be made.

To document this consistency, protocol developers and/or researchers examine a measure's psychometric properties. The psychometric process typically begins in the early stages of development, including considering which constructs will be focused on, examining the scholarly literature on the theoretical basis for these constructs, and meeting with experts to generate and review draft items. Recalling once again the definition of validity, particularly important during this initial process is explicitly linking the purpose of a measure with the claims that are to be made regarding which aspects of classroom quality are being measured, as well as the evidence to back up those claims. After piloting an instrument, assessment developers also may conduct "think-aloud" interviews to determine why observers scored pilot items in a particular way. Parts of this cycle may be repeated as feedback is incorporated into the measure revision process as well.[42]

Additional psychometric research can be conducted once the development phase is complete. This can include field-testing an assessment on a large, representative sample and conducting analyses to determine a measure's validity with the scores from similar measures, as well as its predictive validity (e.g., relationship between classroom quality as measured

by the ECERS-R or CLASS PreK and children's academic outcomes). Also of interest may be the degree to which scores for a classroom are stable over time and independent observers assign similar scores to that classroom, particularly when there otherwise is no reason to expect a change in classroom quality.[43]

The psychometric research on the ECERS-R[44] and CLASS Pre-K[45] is mixed in terms of the quantity and methodological rigor of studies focusing on any of these topics. One indication of the need for even more definitive research is evidenced by a 2013 federal Institute of Education Sciences award to researchers at the University of Illinois. Their study will examine the predictive, content, and structural validity of the ECERS-R and CLASS-PreK when used as part of state QRIS efforts and the Head Start recompetition.[46]

Standardized scoring procedures. Finally, to support its psychometric integrity, a protocol should have standardized procedures for conducting and scoring observations. As an example, the standard approach to scoring the ECERS-R is a "stop-score" process, meaning that to assign a numerical rating of 2–6 to any of the 43 items, an observer must see evidence for all of the indicators for the lower numerical score(s), as well as half of the indicators for the current score (or all of the indicators to score the highest rating of 7). At the same time, some higher-ranked, non-scored indicators may be present in a classroom, and thus missed when following the standard rules. However, such indicators can be acknowledged if the "Alternate Scoring Option" approach is used.

Of course, if two observers rate the same classroom with the ECERS-R, but one uses the stop-score rules and the other the alternate option, their final observation scores may differ.[47] In fact, analysis of ECERS-R data from 265 public school PreK, Head Start, and child care classrooms found that as many as 28 percent of the classrooms that did not meet state quality cut scores using the traditional stop-score approach might meet those cutoffs if scored using an alternate method.[48]

In summary, the validity of observation data is dependent on the extent to which there is sufficient evidence to support the interpretation of a protocol's scores for a particular purpose and for a specific population or setting. It therefore is important to consider the match between the goal of collecting observation data and the protocol used, as well as the measure's psychometric properties. However, while this specific evidence is necessary, it is only one aspect of the validity and reliability equation. As the next section explains, also crucial is evidence of the extent to which an observer can accurately score a protocol.

Observer Capacity to Generate Reliable Scores

A second key policy issue to consider is: Who should conduct classroom observations, and what specialized knowledge base, skills, and oversight, if any, will these individuals need to effectively contribute to the evidence supporting the validity of interpreting score data for a specific purpose and population? The topic of enhancing observer capacity also includes the training needed to understand a protocol; notice the materials, interactions, or teacher practices of interest; and determine the appropriate score for any item based on the protocol's rubric and the evidence gathered.[49] Of additional importance is by what method and how often observers' ongoing scoring reliability will be tracked.

Familiarity with the early education program and/or teachers being observed. When considering who should conduct classroom observations, early education policymakers need to decide if observers should be drawn from "in-house" staff or "outside" consultants with no professional link to individual providers or teachers. The case for in-house staff may be driven by the purpose of a monitoring effort. For example, if observation scores are intended to inform how teachers' practice might be enhanced, using existing curriculum coaches or master teachers as observers provides the potential for any subsequent professional development or technical assistance to be embedded within a continuous feedback cycle.[50] Cost constraints also may play a role in this decision —

e.g., using observers that already are on site or living in a specific geographic area may mean they can conduct observations more frequently and/or less expensively.[51]

Yet, the organizational management literature,[52] research on early education child-focused rater reliability,[53] and principal-as-rater examples from K–12 settings[54] suggest prior professional and personal relationships between raters and the individuals being evaluated may lead to biased scores. For example, similar percentages of Chicago school principals and highly trained observers scored teachers as being in the "unsatisfactory/basic" range based on the protocol used. However, when determining which teachers should be rated as being in the highest category, 17 percent of principals gave teachers this rating versus only 3 percent of the external observers.[55] In a second study, researchers compared the average observation scores assigned to teachers by their own administrators versus administrators from a different school. Same-school administrators tended to score their own teachers more favorably than the other-school administrators. And, when teachers observed other teachers, they were more likely to score their observations in the mid-range, rather than rating a fellow teacher as below basic or above proficient.[56] In summary, even if such bias is not intentional, it must be considered. This is particularly the case if observation score data have consequential implications, such as contributing to official teacher evaluations.

Minimum prior knowledge base and skills. Whether using in-house staff or outside consultants as observers, policymakers also must consider what prior knowledge base and skills observers should have. Because observations will take place in settings serving young children, it might be assumed that the most reliable scores can be generated by individuals already possessing some level of early childhood knowledge and experience. However, at present there is little guidance available regarding the exact level of prior background knowledge necessary to reliably score either of the highlighted protocols, much less how differences in background influence score reliability.

Moreover, the degree to which prior knowledge and skills affect an observer's ability to reliably score individual protocols may be dependent on the *inference level* of a protocol's items. As an example, recall that some observation protocols focus on features such as teacher-child ratios, the square footage of the classroom, or the materials and equipment available (e.g., a sand or water table). Because these programmatic inputs might be determined through simple math or documentation, they can be thought of as "low-inference" items.[57] In short, if an observer can count and use a tape measure, it presumably will be easy for him or her to determine the score for these items based on the rubric being used, even without any prior experience in settings serving young children.

In contrast, other observation protocols examine the nature of teachers' interactions with preschoolers and teaching practices within a specific academic domain, such as the extent to which a teacher supports preschoolers' math and science learning while engaged in activities and discussions at the sand or water table. These types of "higher-inference" items require the observer to accurately recognize behaviors that may be far less obvious to the untrained eye.[58] As might be expected, such items may be more difficult to score, not to mention serve as a source of variation in scores across multiple observers.[59] It therefore may be helpful for these observers to have completed early childhood coursework and/or have experience working in early education classrooms.

At the same time, experience can be detrimental to the production of reliable scores for high-inference items if it results in observers perceiving that they "know" what constitutes good teaching and thus privileging their personal opinions over the measure's scoring rubric.[60] Threats to score reliability also can occur when performance assessment raters elect to use their own reasoning processes for determining an appropriate score.[61] Observers may seek out evidence to confirm their initial, non-rubric related judgment, as

well.[62] Additional research suggests that observer beliefs may play a role in their ability to be deemed reliable.[63] No matter what the source, if an observation protocol contains high-inference items, in addition to observer training and experience, it may be especially important for the scoring rubric to be specific enough to reduce the likelihood that an observer will be able to incorporate or rely on irrelevant ideas regarding what counts as high or low quality.[64]

Training. Since prior knowledge, skills, or program familiarity may not be sufficient for ensuring that observers will accurately use an observation measure's scoring rubric, it perhaps is not surprising that training has been characterized as "one of the most important tools" for ensuring that performance assessment scores will be reliable, particularly if multiple observers will conduct observations across numerous settings.[65] Accordingly, the joint position statement from the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education stresses that "assessment administrators" should be adequately trained and proficient in administering and scoring any measure.[66] Similarly, the National Association for the Education of Young Children and the National Association of Early Childhood Specialists in State Departments of Education advise that program evaluations should be conducted by "well-trained individuals who are able to evaluate programs in fair and unbiased ways."[67]

While training clearly is important, due to the sheer number of early education-focused observation protocols available, no one-size-fits-all model of best practices exists. Instead, training might be led by the protocol developers themselves, someone who has completed an official "train the trainer" series of classes, or PreK in-house staff or consultants.[68] The amount of time required to complete observer training on some of the more well-known early childhood-focused protocols differs, too, ranging from two days to about a week.[69] In addition, the amount of observer training required may be dependent in part on the nature and/or quantity of protocol judgments to be made.[70] For example, training that lasts just a few hours may be suitable for protocols with easily observed, low-inference items. However, when evaluating high-inference classroom practices and interactions, observers may benefit from a more intensive level of training and practice.[71] Training needs may differ depending on observers' prior experience with a protocol as well.[72]

In sum, policies regarding the amount of protocol training needed and who provides it ideally will reflect what is necessary for any group of observers to obtain reliable scores. Given that developer-provided training on the CLASS Pre-K and ECERS-R costs $850 and $1,500 per person, respectively, these policies also must be juxtaposed against program budgets.[73]

Initial and ongoing scoring reliability. Once observers are trained, they need to demonstrate that they can accurately use the protocol's scoring rubric. Recalling the definition of reliability, demonstrating such accuracy helps ensure that the score for any observation is dependent on teachers' practice and/or children's experiences on a particular occasion, rather than on the observer. New observers typically demonstrate their capacity to generate reliable scores by comparing their scores with those of a master coder or the individual leading the training. If the two sets of scores meet or exceed a predetermined percentage of agreement, over a minimum number of observations, the observer is then deemed to be reliable (also known as "calibrated" or "certified").

It might be assumed that once an observer completes his or her training on a protocol and meets the criteria for reliability, it is similar to a person having learned to ride a bicycle in that he or she now has the capacity to generate reliable score data at any point in the future. Unfortunately, initial reliability is not enough to ensure that observers' scores will remain accurate over time. As McClellan et al.[74] note:

> Certification provides assurance that, at the time the assessment was completed, the observer knew the required information and could apply

the rubric at the required level of accuracy and consistency. It does not and cannot guarantee that this observer will always demonstrate this level of performance. People forget, and skills not in continuous use tend to diminish. Even skills in regular use can shift given the context of use.

The tendency for an observer to first be considered certified, but then score more harshly or leniently over time than is warranted by the scoring rubric, is known as "rater drift."[75] As an example, in a study of middle and high school algebra teachers using a version of the CLASS designed for these grades,[76] researchers found that differences in raters' Emotional and Instructional Support scores were dependent on when videos of teachers' practice were scored.[77] Drift also may be more likely when scoring high-inference items as opposed to more clearly defined, low-inference items.[78]

While rater drift (as well as any ongoing differences in rater scoring severity) has implications for the reliability of observation score data, addressing the issue on a large-scale basis also has implications for early education program budget and personnel resources.[79] For example, it can be somewhat easier to detect rater scoring issues when two or three raters focus on the same lesson in the same classroom on the same day, as immediate comparisons of scores can be undertaken and resolved.[80] Having observers meet frequently to discuss issues or receive refresher training also may help; in fact, the CLASS Pre-K Implementation Guide suggests that observers meet at least once a month during any observation cycle so that they can receive feedback about their scores.[81] It also may be useful to "switch up" observation team members so that differences in scoring can be noted more easily.[82]

However, to minimize labor costs or maximize a limited number of observers, policymakers might elect to have just one observer collect data from any one classroom at a time. If so, additional decisions are needed regarding by what means any individual's ongoing scoring competency will be established, such as analyzing differences in his or her scores and/or having a second individual co-score a classroom upon reaching a predetermined time interval or number of observation occasions. Some QRIS initiatives follow this approach in that the number of ongoing drift checks differs based on whether observers are new or more experienced, how many months have passed since their previous reliability check, and their ability to maintain an average percent agreement with a master coder.[83]

Classroom Sampling and Frequency of Observations

As mentioned earlier, periodic administration of any observation measure has the potential to provide policymakers with the opportunity to engage in an informed "plan, do, review, and act" model. However, the degree to which observation data can adequately inform specific decisions about teachers, individual providers, or an early education program overall is dependent on the quantity of reliable data from each of these sources over time. Therefore, in addition to considering the protocol(s) used and capacity of any observer to generate reliable score data on any activities they observe, a third policy issue to be considered is: How many and which classrooms should be observed within any one observation cycle? This policy may be particularly important if the observation scores will inform "high-stakes" decisions regarding individual teacher or provider contracts.

Recalling once again the definition of validity, any decisions regarding whether to observe all or some subset of classrooms during an observation cycle ideally will be driven by the purpose of collecting such data and the evidence needed to support the interpretation of scores for a particular population. For example, if the aim is to determine overall teacher professional development needs, it may be perfectly fine to sample teachers and/or individual providers on a purely random basis, or instead sort and sample them based on such criteria as years of experience or number of children enrolled.[84] If the goal is to produce a summary score that

accurately represents the quality of multiple classrooms within a single center (e.g., as part of a state's QRIS effort), a more deliberate sampling approach may be required, as there can be a greater likelihood of misclassifying a center's quality when randomly sampling just one classroom vs. 50 percent of, or even all, classrooms.[85] This overall research base also suggests that if observation score data are being collected to make judgments about individual teachers' effectiveness and/or whether they should be rehired, optimally the classrooms of these specific teachers will be observed.

A related issue is how frequently any one classroom should be observed within an observation cycle so that score data accurately reflect young children's daily experiences.[86] This issue is important because what a single reliable observer sees on any given day may not be similar to what that same observer — or another reliable observer — would witness if they returned to the same classroom on a different day. The limited research base focused on K–12 classroom rater reliability highlights this issue. For example, analysis of observation data for middle and high school algebra teachers over the course of a school year suggests that their practice is not consistent.[87] Not surprisingly, an analysis of observation score data for 67 elementary, middle, and high school teachers found that the reliability of a single score for determining the quality of any teacher's practice can be low.[88]

This limited research base also illustrates the challenges in establishing a hard-and-fast rule regarding how many observations are needed over the course of a school year to reliably generalize scores to individual teachers. Examination of CLASS score data from classrooms serving older children suggests that observations should take place at least two times per year to generate more reliable scores.[89] Other research on observations of 34 middle school math and English language arts teachers found that reliable scores could be generated with just two observations. However, the protocols used in this particular study had a very narrow scope and just six or 11 items.[90] A study of scores related to administration of a mathematics-focused protocol in middle schools suggests lower score reliability when observers document a single lesson versus three different observers scoring the same teacher on three different lessons.[91] Furthermore, if the results of observations contribute to a high-stakes decision, even four observations may not be enough to produce reliable scores.[92]

In addition to determining the frequency of observations over time, also to be considered is at what time an observation should be scheduled on any single occasion so that observers will be present when the activities or teaching practices focused on in the observation protocol are taking place.[93] Data from Grade 5 classroom observations found that teachers' instruction can vary depending on the academic subject being taught.[94] Even when teaching the same subject, elementary teachers' practice can vary based on the curriculum.[95] The issue of when observations must be scheduled to adequately assess domain-specific teaching may be especially salient in early education classrooms, where the teaching of academic subjects often does not take place in isolation and instead is integrated with a variety of classroom experiences.[96]

Also, what observers see in early education classrooms may depend on whether teachers are leading a small group early literacy lesson versus the whole class being on the playground or engaged in free-choice time, as well as the ages of the children present.[97] The exact time of day may or may not matter: an examination of data collected across a range of state-funded PreK classrooms suggests that teacher-child interactions are relatively stable during the first two hours of the day.[98] However, research in Grade 3 and 5 classrooms suggests that levels of instruction may be lowest during the first 25 minutes of the school day and when students and teachers typically are "settling in."[99]

Of course, when implemented in non-research settings and by early education policymakers, sampling

and frequency decisions most likely will be driven by the supply of trained observers and/or the cost of conducting observations. For example, research on administration of the CLASS Pre-K estimates the direct costs of observation to be $300–$500 per classroom (but also including the cost of initial observer training).[100] There can be significant travel expenditures as well, particularly if a small number of observers are responsible for individual providers located across a wide geographical area. As a result, tradeoffs in the number of classrooms and/or frequency may need to be considered.

In summary, while classroom observation score data have the potential to play an important role in an early education program's monitoring efforts, policymakers must be mindful of the degree to which observation scores provide sufficient evidence to support a decision regarding a teacher, individual provider, or the early education program as a whole. Also to be considered is alignment of the protocol with the monitoring goal(s) and setting, as well as the capacity of the observers to consistently use the protocol and make accurate scoring judgments. All of these concerns must be balanced against monitoring budget and personnel constraints.

Given these concerns, as well as the growing reliance on observation score data for an increasing number of consequential decisions, it would be helpful for the early education field to examine the variety of current classroom observation policies as a means of informing the agenda for continued discussions on best practices. The second purpose of this report is to provide the status of such policies in state-funded PreK. The results of this inquiry are presented next.

To illustrate the saliency of these validity and reliability topics for the early education field's discussion agenda regarding classroom observation best practices, this section provides a description of PreK classroom observation policies for the 2012–2013 school year. Information on these policies was gathered via an author-designed survey of the administrators of the 53 PreK programs identified in NIEER's 2011 *Preschool Yearbook.*[101] The survey aimed to address the following research questions:

1. Which PreK programmatic decisions are informed by external observation score data?
2. Which observation protocols are to be used to generate score data?
3. What affiliation, if any, do observers have with the PreK teachers being observed?
4. What training and ongoing reliability supports do observers receive?
5. How frequently are observations to be conducted in any PreK classroom?

Staff representing 47 PreK programs responded to the survey, for a total response rate of 89 percent. When possible, policy information for the remaining six programs was determined through examination of state RTTT-ELC applications and/or online regulations. These applications and online regulations also were used to prompt requests for clarifications from all of the survey participants. The data then were entered into an SPSS database so that descriptive statistics could be generated.

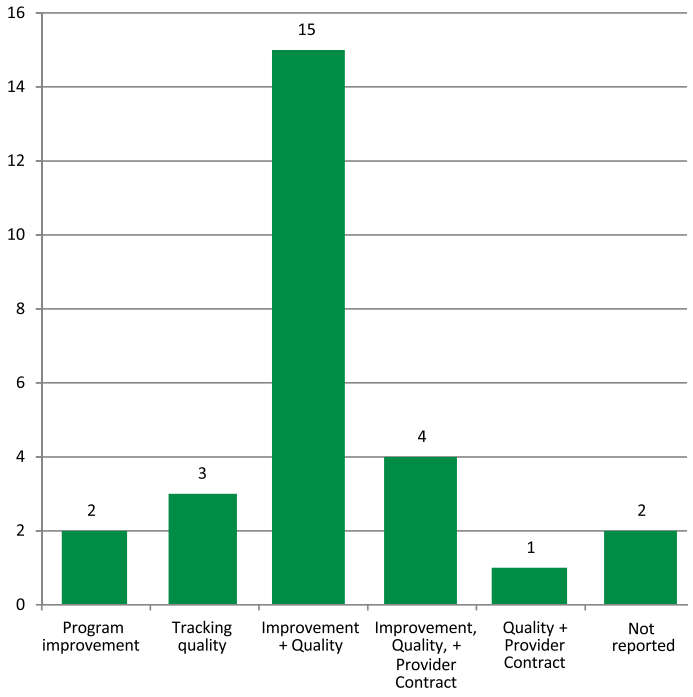
Of the 53 PreK programs, 27 report policies requiring that external observations be conducted. Among the 26 remaining PreK programs, classrooms may be observed as part of the National Association for the Education of Young Children accreditation or Head Start monitoring process, or state-based QRIS efforts. In addition, some PreK programs allow their regional agencies and/or school districts to regulate such policies. PreK programs may require teachers to conduct self-assessments as well. However, these additional policies were not the focus of the study and thus are not included here.

Which PreK Programmatic Decisions Are Informed by External Observation Score Data?

The study's first research question was addressed by asking survey participants to indicate for what purpose(s) observation score data are used. The selected responses for this question included contract or funding level decisions, as well as informing professional development and technical assistance. Respondents also could indicate "other" and provide further details.

Figure 1 displays the reported uses of observation score data. As can be seen, two programs indicate these data are used solely to inform decisions regarding program improvement. This includes teacher professional development and technical assistance, as well as material and equipment needs. Another three PreK programs report observation scores are used solely for the purpose of tracking quality. This includes verifying eligibility for a QRIS rating or center accreditation by the National Association for the Education of Young Children.

Figure 1. *Decisions Informed by Classroom Observation Score Data*



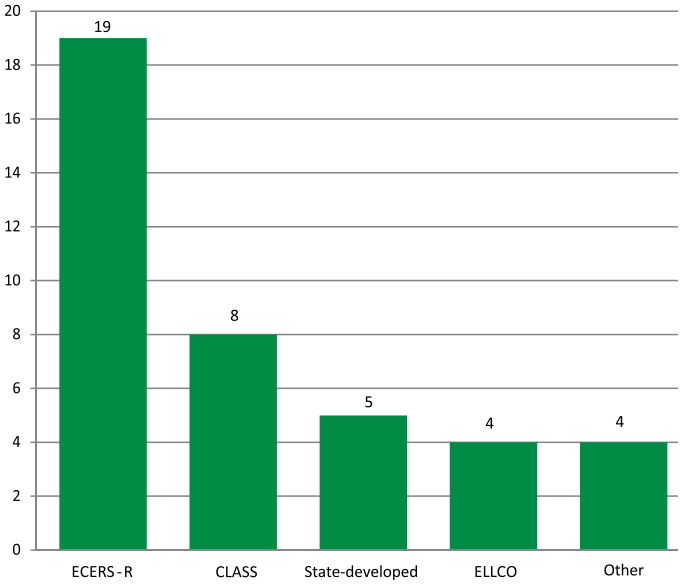
Among the 22 remaining PreK programs, 15 report that score data informs both program improvement and quality tracking efforts. An additional four programs use score data for these purposes as well as to inform decisions about individual provider contracts. One additional program relies on score data for tracking quality and determining whether an individual provider's contract should be issued or renewed. When combining all of these results, 21 programs report observation score data inform program improvement decisions. A total of 23 programs use the data to track classroom quality. It also should be noted that in contrast to K–12 policies throughout much of the United States, no PreK program reports that classroom observation scores are used to evaluate individual teachers. Similarly, no program indicates that scores are used to determine if a teacher's employment contract should be renewed.

Which Observation Protocols Are to Be Used to Generate Score Data?

The study's second research question was addressed by asking survey participants to indicate which classroom observation protocol(s) must or could be used as part of their PreK monitoring process. Given the long-standing use of the ECERS-R in research studies and QRIS initiatives, as well as Head Start reliance on the CLASS Pre-K, the selected responses for this question included these measures. Respondents also could indicate the names of any additional protocols through an "other" response.

Figure 2 displays the total number of PreK programs reporting policies that require the use of specific measures either alone or in combination. As can be seen, 19 of the 27 PreK programs indicate that their program policies require the ECERS-R to be used. This includes use as the sole instrument or with other observation measures. The second most-reported protocol is the CLASS Pre-K, which is required by a total of eight PreK programs.

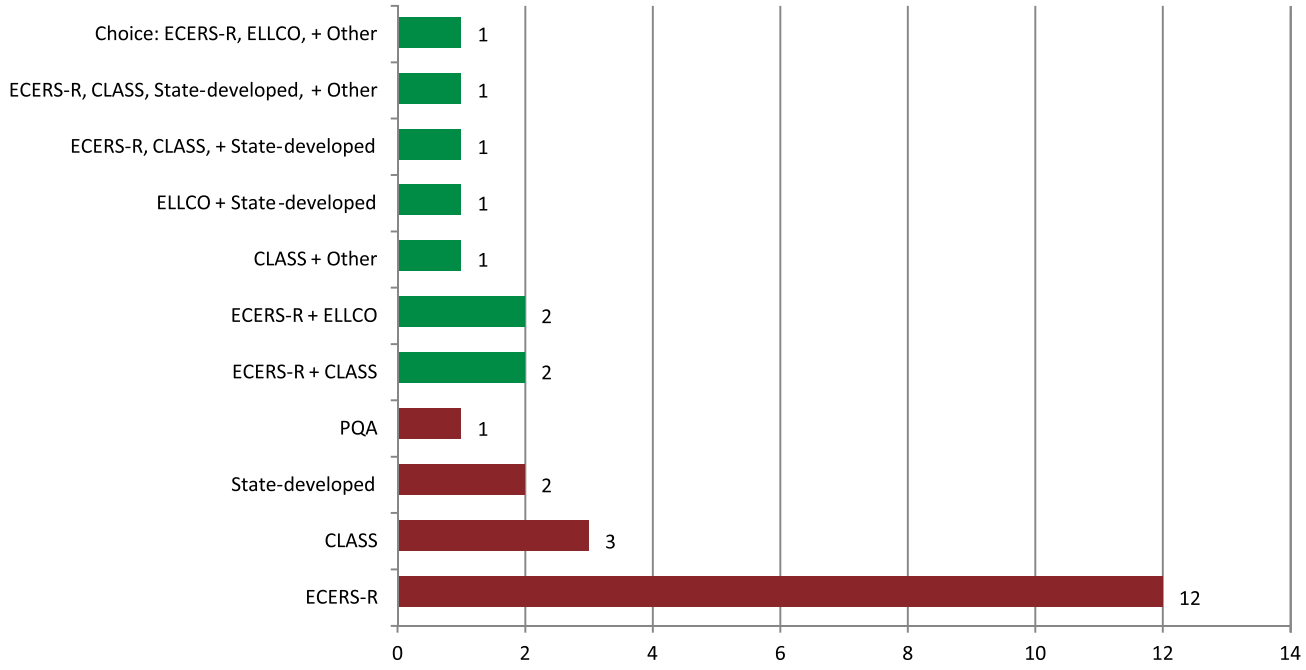
Figure 2. *Number of PreK Programs Mandating Use of Specific Observation Protocols*



An additional five PreK programs report reliance on a state-developed observation protocol. Four PreK programs report using the *Early Language and Literacy Classroom Observation Tool (ELLCO) Pre-K*,[102] which measures classroom support for children's language and literacy development. The protocol has 18 items within the five domains of classroom structure, curriculum, the language environment, books and reading, and print and writing. Finally, four PreK programs report the use of observation protocols that do not align with any of the previous four categories.

Figure 3 reports this same information, but instead of tallying the results for each specific observation protocol, it displays the individual PreK program protocol policies. In this figure, the seven green bars represent the nine PreK programs reporting more than one classroom observation protocol as to be used. The four red bars represent the 18 programs reporting use of a single observation measure.

Figure 3. *Specific Observation Protocol(s) To Be Used*



As can be seen by viewing the green bars, among the nine programs using a combination of protocols, no single specific model emerges. However, seven of these nine programs report that their protocol combination includes the ECERS-R. Five of these programs report that the CLASS Pre-K is used in combination with at least one additional protocol. Four programs use the ELLCO as part of a combination of observation protocols. Finally, just one PreK program reports that individual providers may choose from among a variety of protocols when being observed.

Among the "single measure" policy models (displayed in the four red bars), 12 programs report exclusive use of the ECERS-R. Three programs report policies requiring use of the CLASS Pre-K and two programs report use of a state-developed measure. Finally, one program reports reliance on the *Preschool Program Quality Assessment (PQA)*, which has 63 items within seven domains, including curriculum planning and assessment, parent involvement, and family services.[103]

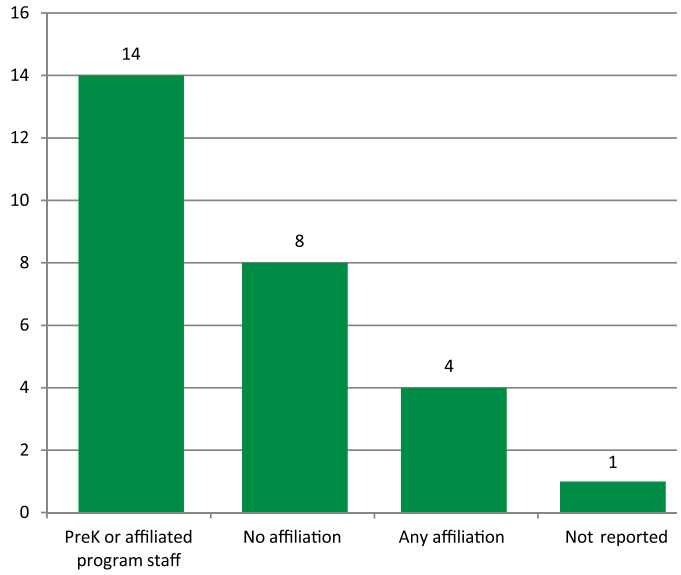
Observer Affiliation and Qualifications

Another set of survey questions focused on the study's third research question, which aimed to determine the extent to which observers have a pre-existing affiliation with the teachers being observed or the individual providers in which teachers' classrooms are situated. To address this research question, the survey asked whether the observers were employed by the state-level department overseeing a state PreK program, a contracting or regional agency that also is responsible for some aspect of administering the PreK program, or an individual provider such as a school district, child care center, or Head Start grantee. If a PreK program relies on outside consultants hired specifically to conduct observations, also of interest was whether these individuals must have a degree related to early childhood.

As can be seen in Figure 4, 14 survey respondents report their observers are employed in some additional capacity within the PreK program itself or an affiliated agency. Conversely, eight PreK programs report classroom observers are outside consultants who are

hired specifically to conduct observations. Four PreK programs report that observers are not restricted to either category. Instead, observers may be already employed in some capacity or hired on a consultant basis solely to perform observations.

Figure 4. *Affiliation of Observers*



Among the 18 PreK programs relying on already employed staff to conduct observations, in 14 cases this employment was within the state-level department overseeing the program. The remaining programs within this group use observers who are employed by a regional agency, county coalition, school district, or other state-level agency that collaborates with the PreK program. For the 12 programs using outside consultants as observers, seven require observers to have a minimum of a B.A. related to early childhood.

It should be noted that the survey did not ask respondents to indicate the motivation behind the choice of observer affiliation. However, anecdotal information gathered during administration of the survey or follow-up telephone conversations suggests a wide continuum of reasons. For those programs using state-, regional- or provider-related staff, these reasons include budget constraints within PreK programs (and thus the lack of capacity to hire consultants) or the intention to have any follow-on professional development be provided by the observers. One of the PreK programs relying on consultants remarked that this choice was an intentional attempt to ensure observers did not have a supervisory relationship with the teachers being observed.

Observer Training and Ongoing Reliability Methods

The fourth research question focused on the training and ongoing scoring reliability supervision provided to observers. The survey therefore also queried respondents about the observer training and certification process, as well as how frequently observer drift checks take place. Twenty-two of the 27 PreK programs responded to these survey questions.

All 22 programs report that observers receive in-person training, with the training often supplemented by the use of videos, reading materials, and/or web-based resources. In addition, 19 programs report that all observers undergo the same training. In the remaining three programs, the training received depends on whether the individual observer also is a trainer of other observers (and thus participates in "train the trainer" classes), his or her prior experience using the protocol, or preference for more training versus a greater number of initial reliability sessions.

Twenty-one of the 22 PreK programs report that observers practice score live and/or videotaped classrooms to help determine their initial scoring accuracy. In addition, 19 of the 21 programs report that before observers are allowed to conduct consequential observations, they are required to produce scores that agree with expert ratings on some predetermined basis. The required number of times an observer needs to favorably score an observation to determine his or her initial reliability varies from zero to five (with just one PreK program reporting no initial reliability is determined). Six programs report the number of times necessary to reach the expected rater-expert agreement rate is based on the recommendations of the protocol's developers.

Seventeen of the 22 programs report that observers undergo ongoing score drift checks, but four additional programs report that no drift checks take place. For the 17 programs that do undertake some type of drift checking, the timing ranges from every 90 days to three years and can vary based on whether an individual is a new or more experienced observer. However, 13 of the 17 programs schedule their observer drift checks at least annually.

Frequency of Observations in PreK Classrooms

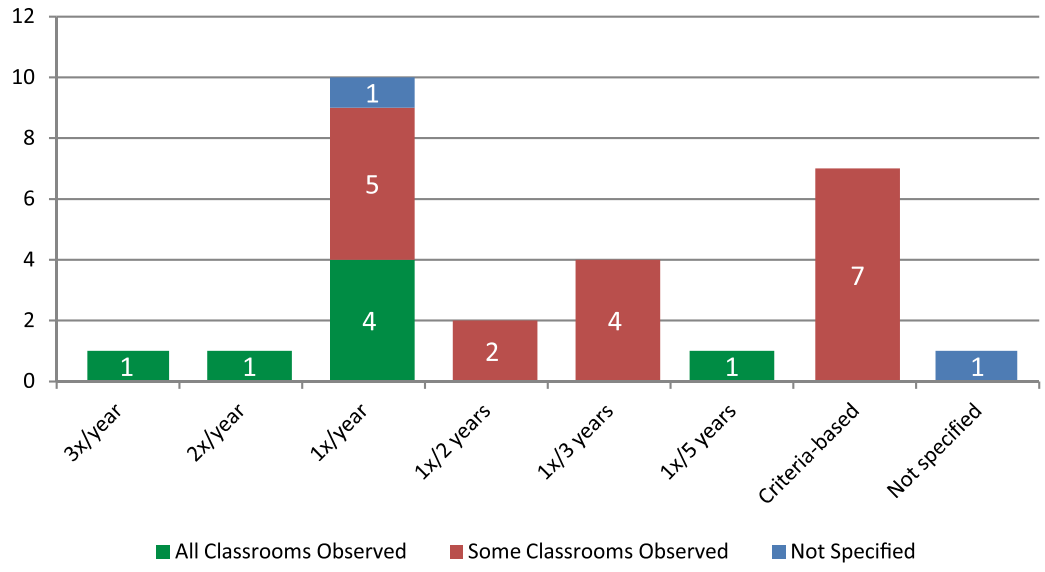
The final research question focused on how frequently observations are to occur within any PreK classroom. For this question, frequency referred to the number of times within an observation cycle (e.g., twice per year, once per year, once every two years), as well as whether all or some classrooms are observed within the cycle. Figure 5 displays the results of this survey question.

Number of observations within a cycle. As can be seen by looking at the column titles on the horizontal axis, the number of observations conducted within a cycle across PreK programs ranges from three times per year to once every five years. However, just two programs observe three or two times per year and 10 programs report observations take place annually. In seven PreK programs, observations take place just once every two, three, or five years.

An additional seven PreK programs report the frequency of observations is criteria-based, rather than part of a predetermined cycle. These criteria include how long an individual provider has participated in the PreK program, whether their contract is under consideration, and their current QRIS rating. Observations also may be dependent on whether a classroom was observed previously or the observation scores were obtained through a teacher self-assessment.

All vs. some classrooms. While these data represent one aspect of the frequency of observations, also important is the likelihood that a specific classroom will be observed within the PreK program's observation cycle. Figure 5 therefore also displays the "all vs. some classrooms" variation within each cycle frequency, with the green bars representing the programs reporting all classrooms are observed and those programs that observe a subset of classrooms shaded in red. As can be seen by looking at

Figure 5. *Observation Frequency and All vs. Some Classrooms Observed*



the green columns, just seven PreK programs report that all classrooms are observed within any cycle. For example, the sole program that observes classrooms three times per year reports doing so in all of its classrooms. Also included in this category are four of the 10 programs reporting that observations take place once per year, as well as the PreK program using a five-year cycle.

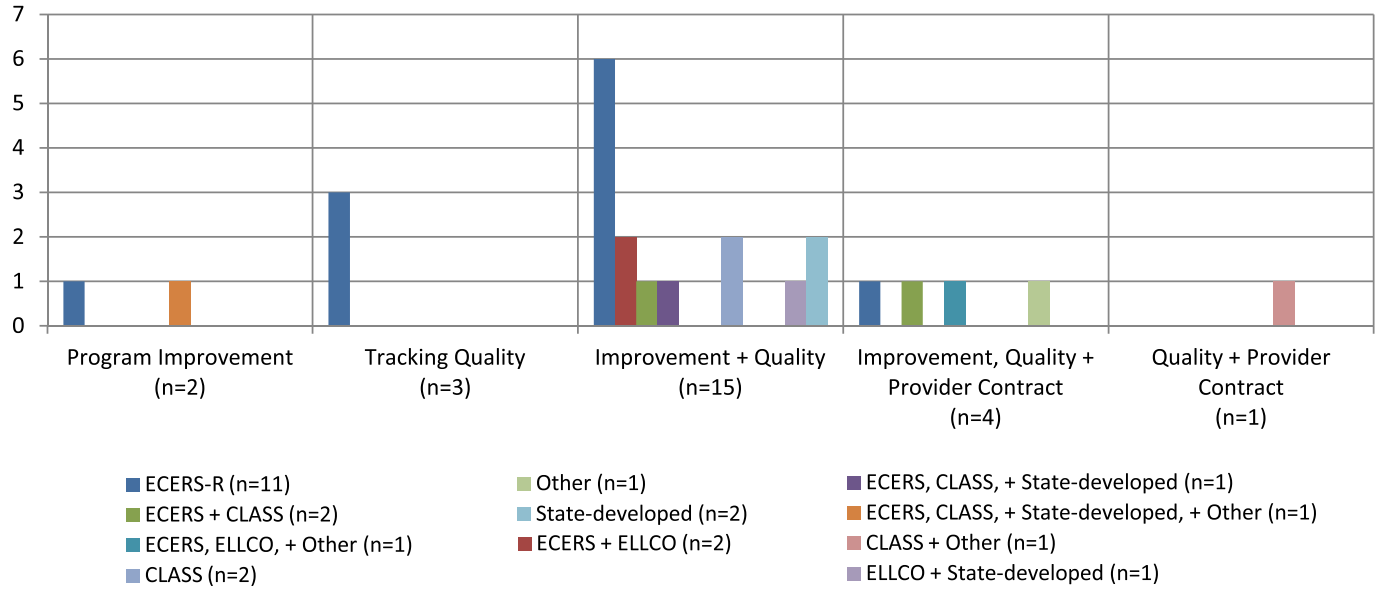
The remaining programs (indicated by the red columns) observe in a subset of classrooms. This includes five of the "once per year" programs and all six programs that observe classrooms every two or three years. Not surprisingly, only a subset of classrooms are observed in the seven PreK programs that use specific criteria to determine whether an observation will be scheduled.

Evidence of PreK Classroom Observation Policy Model Consensus

One final area of interest is whether there appears to be any relationship between the individual PreK survey responses that also might suggest evidence of a common policy model. Of course, the sample for this study is too small to conduct any type of sophisticated statistical analyses. Nonetheless, the survey data were examined overall to determine if there appear to be any correlations for one or more of the policy issues highlighted in this report.

Given that validity is related to the extent to which an interpretation of observation scores is appropriate for any specific task, the first topic investigated was whether there is a correlation across PreK programs between the purposes for which observation data are used and the exact observation protocol models reported. Figure 6 shows the results of this inquiry. In this figure, the same score data purposes displayed in Figure 1 are on the horizontal axis. In addition, the observation protocol models displayed in Figure 3 are represented by the different colored bars within each purpose column. Finally, the quantity of PreK programs using any specific observation model is represented by the numbers on the vertical axis.

Figure 6. *Relationship between Score Data Use and Observation Protocols Used*



As can be seen by looking at Figure 6, the only purpose shared by two or more PreK programs and also having a single protocol model is for "tracking quality" (second column from the left). In this case, all three PreK programs that use observation score data solely to track quality also rely on the ECERS-R. In contrast, in the two PreK programs that use observation score data for program improvement (left-most column), one program relies solely on the ECERS-R, but the second uses the ECERS-R, CLASS, a state-developed measure, and a fourth protocol.

There is even more variation within the 15 programs that rely on observation score data to inform program improvement and tracking quality decisions (middle column). For these dual purposes, a total of seven different protocols models are used. These models range from the ECERS-R alone (n=6), the ECERS and the ELLCO (n=2), the ECERS and the CLASS (n=1), only the CLASS (n=2), and stand-alone, state-developed measures (n=2).

Further examination of the survey responses (see Appendix) shows that how score data are used does not appear to be related to decisions regarding the affiliation of the observers or how frequently any classroom might be observed, much less whether all or some classrooms are part of an observation cycle. To summarize, while this study was a preliminary examination of state PreK classroom observation policies and thus limited in its implications, it suggests that the PreK field as a whole does not appear to have coalesced around a single overall policy model for regulating the collection of external classroom observation score data. Discussed next are the implications of these findings for which validity- and reliability-related topics potentially should be on the agenda for future discussions regarding best classroom observation practices.

This report summarizes 2012–2013 PreK classroom observation policies as a means for highlighting why it might be useful to include a set of key validity- and reliability-related topics in the early education field's ongoing best practices discussion agenda. Such an agenda focus especially may be relevant given the increasing use of early education classroom observation data to inform decisions that can have ramifications for individual teachers, schools, grantees, and overall programs, as well as federal plans to expand 4-year-olds' access to high-quality, publicly funded preschool.

As was highlighted above, just 27 state-funded PreK programs report they require external classroom observation score data to be collected as part of their monitoring process. Twenty-one of the 27 PreK programs rely on observation score data to inform decisions related to program improvement, such as technical assistance or teacher professional development needs. Another common purpose across the majority of PreK programs (n=23) is to track individual provider quality, either more broadly or to contribute to decisions related to QRIS ratings. Both purposes are aligned with the overall focus in the early education field on improving quality. Conversely, the emphasis on high-stakes teacher evaluations that is prevalent in K–12 classrooms has not yet extended to state-funded PreK programs.

While the overall PreK field appears to use classroom observation data for program improvement and/or quality tracking, the results of this study suggest that no single overall policy model exists in terms of the protocol used, affiliation of the observers and the degree to which their reliability is initially determined and then tracked, and the frequency with which observations are conducted. The lack of consensus regarding a policy model may not necessarily be of concern, as PreK programs ideally will craft policies that are aligned with their respective monitoring needs and program resources. However, in light of the validity and reliability issues summarized in the first part of this report, the wide variation in policies also suggests some issues may be worth inclusion on the classroom observation discussion agenda as the field continues to define best practices.

Which Protocol(s) Should Be Used?

Although there are a variety of individual classroom observation protocols available for use in settings serving young children, this study suggests that the ECERS-R is the most common policy-prescribed protocol, either alone or in combination with other measures, in state-funded PreK programs. Because the survey did not ask why any particular protocol was chosen, it is not clear if the overwhelming reliance on the ECERS-R is a result of its traditional status within the early care and education field, the emphasis on QRIS scores within a PreK program, and/or the degree to which early education policymakers, observers, and classroom staff are familiar with the ECERS-R focus and administration. Also unclear is whether the inclusion of the CLASS in a more limited number of PreK program policies is due to an awareness of the validity limitations of using interpretations of ECERS-R data for certain monitoring decisions, and/or the benefits of using more than one measure or a protocol that focuses on a different view of classroom quality.

No matter what the motivation for choosing to use any observation protocol, of potential interest for the best practices discussion is the extent to which the interpretation of score data from each of these well-known measures is valid for a variety of high-stakes purposes, including consequential decisions about individual teachers and PreK providers. In light of federal efforts to expand PreK, a related topic is how to manage the implications of classrooms being considered high-quality within one state's early education program, but not as favorably rated within another state's program due to the different protocols used. Also of interest is whether the lack of consistency in how quality is defined across publicly funded programs matters for children's early learning outcomes.

What Issues Potentially Affect Observers' Capacity to Generate Reliable Score Data for Any Monitoring Purpose?

As highlighted above, 18 of the 27 programs rely on PreK or other program-affiliated employees to conduct classroom observations. It is unclear whether this finding is related to staffing convenience/constraints, available funding, or the emphasis on using score data to inform program improvement efforts, including teacher professional development and technical assistance to individual providers. As the consequential implications of monitoring decisions for early education stakeholders expands, another best practices discussion topic may be under what circumstances programs benefit, yet perhaps also "lose out," when observers are familiar with classrooms, individual providers, and/or the overall program. Such a topic might be informed by the experiences of the four PreK programs that rely on observers who both are — and are not — otherwise affiliated.

Furthermore, two of these 18 PreK programs report that observation score data are used to inform decisions regarding individual provider contracts, and those data are collected solely by program-affiliated employees. Therefore, of additional interest for the best practices agenda may be the rigor with which the reliability of these observers is tracked, as well as how frequently the classrooms in any "potentially defunded" individual provider are observed before decisions are made regarding their PreK contracts.

No matter what the affiliation of observers, one more potential item for the best practices discussion is the handful of programs reporting that no drift checks take place to ensure observers' ongoing scoring reliability. While different definitions of quality already may exist across programs due to the variety of classroom observation protocol models in use, it especially may be difficult to "bank on" the quality of programs in states where ongoing observer reliability is tracked on an infrequent basis or not at all.

How Frequently Should Observations Take Place?

Currently, 24 of the 27 PreK programs report that they conduct observations in a subset of classrooms and no more than once per year (or even less frequently). The majority of PreK programs also report using observation score data for what might be considered "lower-stakes" purposes, including program improvement. As a result, such an observation schedule may not only be more program resource friendly, but also be appropriate for generating sufficient data to make this type of monitoring decision.

However, if early education programs use observation score data to make high-stakes provider funding or contract decisions, assuming that ongoing observer reliability also is monitored, it will be imperative for future best practices discussions to investigate the number of classrooms, as well as the number of observing occasions within any classroom, that are sufficient for generating reliable score data. This especially will be critical if these data also inform decisions regarding the effectiveness of early education teachers.

Conclusion

Collecting valid and reliable monitoring data likely will remain a top priority as states and the federal government continue to expand children's access to early education programs. This particularly is the case due to concurrent pressure to improve and ensure the quality of these programs and prove they are worth their cost,[104] as well as the larger K–12 policy focus on the measurement of teacher effectiveness. Generating reliable classroom observation score data as part of this monitoring poses some unique challenges. Yet, because such data can provide critical information for both program improvement and accountability decisions, it is imperative that PreK stakeholders can bank on the accuracy of the data and its interpretation for any specific purpose and population. Given the less-than-robust early childhood literature base on the potential validity and reliability issues related to policies on classroom observation protocols, observer capacity, and frequency of observation data collection, the time may be right for early education stakeholders to include such topics in their best practices agenda.

Appendix

	Observation Protocol to Be Used					Observation Score Data Uses			Observer Affiliation		Potential Generalizability of Score Data	
PreK Program Reporting Observation Policies (n=27)	**ECERS-R**	**ELLCO**	**CLASS Pre-K**	**State-developed**	**Other**	**Program Improvement[105]**	**Track Quality[106]**	**Provider Contract**	**PreK or other affiliated program staff**	**No affiliation**	**Frequency of observations in any classroom**	**All vs. some classrooms**
Alabama First Class Voluntary Pre-Kindergarten Program	x	x				x	x		x		1x/year	All
Alaska Prekindergarten Program	x		x			x	x		x		1x/year	All
Arkansas Better Chance/Arkansas Better Chance for School Success	x					x	x	x		x	1x/2 years	Some
California State Preschool Program	x					x			x		1x/year	Some
Connecticut School Readiness	x						x			x	1x/year	Some
District of Columbia Public Charter School Pre-Kindergarten			x		x		x	x	x	x	Based on charter renewal/review status or low performance indicators	Some
District of Columbia Public School Pre-kindergarten (DCPS & CBOs)			x			Unknown			x		1x/year	Unknown
Georgia Pre-K Program			x			x	x		x	x	1x/year	Some
Illinois Preschool for All	x					Unknown			Unknown		Unknown	Unknown
Iowa Statewide Voluntary Preschool Program				x		x	x		x		Based on years in program	Some
Kentucky Preschool Program	x					x	x		x		1x/5 years	All
Louisiana Cecil J. Picard LA4 Early Childhood Program	x		x	x		x	x			x	As needed; no more than 1x/year	Some
Louisiana Non-Public Schools Early Childhood Development Program		x		x		x	x		x		2x/year	All
Massachusetts Universal Pre-Kindergarten and Grant 391 Program	x						x			x	Based on self-assessed QRIS level	Some
Michigan Great Start Readiness Program					x	x	x	x	x	x	3x/year	All
Nebraska Early Childhood Education Program	x					x	x		x		Based on year of PreK grant	Some
Nevada State Prekindergarten Education Program	x	x				x	x			x	1x/year	Some
New Jersey Former Abbott and Expansion Districts[107]	x	x			x	x	x	x	x		1x/year	All
New Mexico PreK	x					x	x		x		1x/year	Some
North Carolina NC PreK	x						x		x		1x/3 years	Some
Ohio Early Childhood Education				x		x	x		x		Based on QRIS step	Some
Oregon Head Start Prekindergarten			x			x	x		x		1x/3 years	Some
Pennsylvania Pre-K Counts	x					x	x			x	1x/2 years	Some
Rhode Island Prekindergarten Program	x		x			x	x	x	x		Based on protocol	Some
Vermont Early Education Initiative	x					x	x			x	1x/3 years	Some
Vermont Prekindergarten Education – Act 62	x					x	x			x	1x/3 years	Some
West Virginia Universal Pre-K[108]	x		x	x	x	x			x	x	1x/year	All

Endnotes

[1] Frede, E. C., Gilliam, W. S., & Schweinhart, L. J. (2011). Assessing accountability and ensuring continuous program improvement. In E. Zigler, W. S. Gilliam, & W. S. Barnett (Eds.), *The pre-k debates: Current controversies & issues*. Baltimore: Paul H. Brookes Publishing Co. Riley-Ayers, S., Frede, E., Barnett, W. S., & Brenneman, K. (2011). *Improving early education programs through data-based decision making*. New Brunswick, NJ: NIEER.

[2] Barnett, W. S., Carolan, M. E., Fitzgerald, J., & Squires, J. H. (2012). *The state of preschool 2012: State preschool yearbook*. New Brunswick, NJ: NIEER.

[3] Head Start Program Performance Standards, 45 CFR Parts 1307.3. (2011). *Policies and procedures for designation renewal of Head Start and Early Head Start grantees.* Retrieved November 18, 2011 from http://eclkc.ohs.acf.hhs.gov/hslc/Head%20Start%20Program/Program%20Design%20and%20Management/Head%20Start%20Requirements/Head%20Start%20Requirements/1307.

[4] US Department of Education (DOE) and US Department of Health and Human Services (DHHS). (2011). *Race to the Top – Early Learning Challenge*. Washington, DC: Author.

[5] Stoney, L. (2012). *Unlocking the potential of QRIS: Trends and opportunities in the Race to the Top – Early Learning Challenge applications*. Retrieved February 12, 2013 from http://www.qrisnetwork.org/sites/all/files/resources/gscobb/2012-03-07%2008:29/LouiseStoneyMemo.pdf.

[6] Guss, S. S., Norris, D. J., Horm, D. M., Monroe, L. A., & Wolfe, V. (2013). Lessons learned about data utilization from classroom observations. *Early Education and Development, 24*, 4–18.

[7] Gitomer, D. H. (Ed.) (2009). *Measurement issues and assessment for teaching quality*. Thousand Oaks, CA: Sage. Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness.* Washington, DC: National Comprehensive Center for Teacher Quality. National Council on Teacher Quality. (2012). *State of the states 2012: Teacher effectiveness policies.* Washington, DC: Author.

[8] Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (In press). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*.

[9] Leeuw, F., & Vaessen, J. (2009). *Impact evaluations and development: NONIE guidance on impact evaluation*. Washington, DC: The Network of Networks on Impact Evaluation.

[10] Barnett et al. (2012).

[11] See http://www.whitehouse.gov/the-press-office/2013/02/13/fact-sheet-president-obama-s-plan-early-education-all-americans for the full plan.

[12] This report focuses on PreK policies governing classroom observations that are conducted by someone other than the teacher herself (or what instead would be referred to as a self-assessment).

[13] See http://www.whitehouse.gov/the-press-office/2013/02/13/fact-sheet-president-obama-s-plan-early-education-all-americans for the full plan.

[14] Barnett, W. S., Robin, K. B., Hustedt, J. T., & Schulman, K. L. (2003). *The state of preschool: 2003 State preschool yearbook*. New Brunswick, NJ: NIEER.

[15] Barnett et al. (2012).

[16] Research and Policy Committee of the Committee for Economic Development. (2006). *The economic promise of investing in high-quality preschool: Using early education to improve economic growth and the fiscal sustainability of states and the nation.* Washington, DC: Author. Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., Ludwig, J., Magnuson, K. A., Phillips, D., & Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education.* New York: Foundation for Child Development.

[17] Ackerman, D. J., Barnett, W. S., Hawkinson, L. E., Brown, K., & McGonigle, E. A. (2009). *Providing preschool education for all 4-year-olds: Lessons from six state journeys*. New Brunswick, NJ: NIEER. Hustedt, J. T., & Barnett, W. S. (2011). Private providers in state Pre-K: Vital partners. *Young Children, November*, 42–48.

[18] Deming, W. E. (2000). *Out of the crisis*. Cambridge, MA: MIT Press.

[19] National Association for the Education of Young Children and the National Association of Early Childhood Specialists in State Departments of Education. (2003). *Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8*. Washington, DC: Author. National Early Childhood Accountability Task Force. (2007). *Taking stock: Assessing and improving early childhood learning and program quality*. New York: Foundation for Child Development. Kochanoff, A., Hirsh-Pasek, K., Newcombe, N., & Weinraub, M. (2003). *Using science to inform preschool assessment*. Philadelphia: Center for Improving Resources in Children's Lives, Temple University.

[20] Frede et al. (2011). Riley-Ayers et al. (2011).

[21] Barnett, W. S., Carolan, M. E., Fitzgerald, J., & Squires, J. H. (2011). *The state of preschool 2011: State preschool yearbook*. New Brunswick, NJ: NIEER.

[22] Ackerman, D. J., & Coley, R. (2011). *State pre-K child assessment policy: Issues and status*. Princeton: ETS.

[23] Barnett et al. (2011).

[24] Harms, T., Clifford, R. M., & Cryer, D. (2005). *Early Childhood Environment Rating Scale – Revised Edition.* New York, NY: Teachers College Press.

[25] Harms, T., & Clifford, R. M. (1980). *The Early Childhood Environment Rating Scale*. New York: Teachers College Press.

[26] Brassard, M. R., & Boehm, A. E. (2007). *Preschool assessment: Principles and practices*. New York: The Guildford Press. Clifford, R. M., Reszka, S. S., & Rossbach, H. (2010). *Reliability and validity of the early childhood environment rating scale (working paper)*. Chapel Hill, NC: University of North Carolina. Hamre, B. K., & Maxwell, K. L. (2011). *Best practices for conducting program observations as part of quality rating and improvement systems*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services. LoParo, K. M., Thomason, A. C., Lower, J. K., Kintner-Duffy, V. L., & Cassidy, D. J. (2012). Examining the definition and measurement of quality in early childhood education: A review of studies using the ECERS-R from 2003 to 2010. *Early Childhood Research & Practice, 14*, online journal retrieved July 9, 2012 from http://ecrp.uiuc.edu/v14n1/laparo.html. Malone, L., Kirby, G., Caronongan, P., Tout, K., & Boller, K. (2011). *Measuring quality across three child care quality rating and improvement systems: Findings from secondary analyses (OPRE Report 2011-30)*. Washington, DC: US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. Pianta, R. C. (2012). *Implementing observation protocols: Lessons for K–12 education from the field of early childhood*. Washington, DC: Center for American Progress. 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*. Washington, DC: Child Trends. Zellman, G. L., Perlman, M., Le, V., & Setodji, C. M. (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.

[27] Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System*: *Pre-K*. Baltimore: Paul H. Brookes Publishing Co.

[28] Head Start Program Performance Standards, 45 CFR Parts 1307.3. (2011).

[29] Joint Committee on Standards for Educational and Psychological Testing. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, The American Psychological Association, and the National Council on Measurement in Education. Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162.

[30] Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational Researcher,* 38(7), 518–521. Frechtling, J., Mark, M. M., Rog, D. J., Thomas, V., Frierson, H., Hood, S., & Hughes, G. (2010). *The 2010 use-friendly handbook for project evaluation*. Washington, DC: NSF.

[31] Leeuw, F., & Vaessen, J. (2009).

[32] Bryant, D. (2010). *Observational measures of quality in center-based early care and education programs* (Research-to-policy, research-to-practice brief OPRE 2011-10c). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Canadian Centre for Knowledge Mobilisation. (2006). *CCKM's research guide for child care decision making.* Retrieved September 28, 2011 from http://www.cckm.ca/ChildCare/ToolsQuality.htm. Grinder, E. L. (2007). *Review of early childhood classroom observation measures*. Harrisburg, PA: Early Learning Standards Task Force and Kindergarten Assessment Work Group, Pennsylvania BUILD Initiative, Pennsylvania's Departments of Education and Public Welfare. Halle, T., & Vick, J. (2007). *Quality in early childhood care and education settings: A compendium of measures*. Washington, DC: Author.

[33] Lambert, R. G. (2003). Considering purpose and intended use when making evaluations of assessments: A response to Dickinson. *Educational Researcher, 32*(4), 23–26. Martinez-Beck, I. (2011). Why strengthening the measurement of quality in early childhood settings has taken on new importance. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. xviii–xxiv). Baltimore: Brookes Publishing. Snow & Van Hemel. (2008). Stuhlman, M. W., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2010). *A practitioner's guide to conducting classroom observations: What the research tells us about choosing and using observational systems*. Charlottesville, VA: University of Virginia. Zaslow, M., Tout, K., & Halle, T. (2011). Differing purposes for measuring quality in early childhood settings. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. 389–410). Baltimore: Brookes Publishing.

[34] Brenneman, K. (2011). Assessment for preschool science learning and learning environments. *Early Childhood Research & Practice, 13*(1). Online journal available at http://ecrp.uiuc.edu/v13n1/brenneman.html.

[35] Guernsey, L., & Ochshorn, S. (2011). *Watching teachers work: Using observation tools to promote effective teaching in the early years and early grades.* Washington, DC: New America Foundation.

[36] Castro, D. C., Espinosa, L. M., & Paez, M. M. (2011). Defining and measuring quality in early childhood practices that promote dual language learners' development and learning. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. 257–280). Baltimore: Brookes Publishing. Halle, T., Martinez-Beck, I., Forry, N. D., & McSwiggan, M. (2011). Setting the context for a discussion of quality measures. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. 3–10). Baltimore: Brookes Publishing. Lambert, M. C., Williams, S. G., Morrison, J. W., Samms-Vaughan, M. E., Mayfield, W. A., & Thornburg, K. R. (2008). Are the indicators for the language and reasoning subscale of the Early Childhood Environment Rating Scale-Revised psychometrically appropriate for Caribbean classrooms? *International Journal of Early Years Education, 16*, 41–60. Shivers, E. M., Sanders, K., & Westbrook, T. R. (2011). Measuring culturally responsive early care and education. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. 191–225). Baltimore: Brookes Publishing. Spiker, D., Hebbeler, K. M., & Barton, L. R. (2011). Measuring quality of ECE programs for children with disabilities. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. 229–256). Baltimore: Brookes Publishing.

[37] Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers., S. S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*, 529–542. Stuhlman et al. (2010).

[38] Bryant, D. M., Burchinal, M., & Zaslow, M. (2011). Empirical approaches to strengthening the measurement of quality. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. 33–47). Baltimore: Brookes Publishing. Layzer, J. I., & Goodson, B. D. (2006). The "quality" of early care and education settings: Definitional and measurement issues. *Evaluation Review, 30*, 556–576. Hill, H. C., Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*, 371–384.

[39] LoParo et al. (2012).

[40] Brassard & Boehm. (2007). Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the Quality of Pre-Kindergarten Teacher-Child Interactions and Instruction. *Applied Developmental Science, 12*, 140–153. Denny, J. H., Hallam, R., & Homer, K. (2012). A multi-instrument examination of preschool classroom quality and the relationship between program, classroom, and teacher characteristics. *Early Education & Development, 23*, 678–696. Dickinson, D. K. (2002). Shifting images of developmentally appropriate practice as seen through different lenses. *Educational Researcher, 31*, 25–32. Dickinson, D. K. (2006). Toward a toolkit approach to describing classroom quality. *Early Education & Development, 17*, 177–202. Hallam, R., Fouts, H., Bargreen, K., & Caudle, L. (2009). Quality from a toddler's perspective: A bottom-up examination of classroom experiences. *Early Childhood Research & Practice, 11*(2). Online journal available at http://ecrp.uiuc.edu/v11n2/hallam.html. Katz, L. (1993). *Five perspectives on quality in early childhood programs (Catalog #208)*. Champaign, IL: Early Childhood and Parenting Collaborative, College of Education, University of Illinois at Urbana-Champaign. Available at http://ecap.crc.illinois.edu/eecearchive/books/fivepers.html#intro. Ross, C., Moiduddin, E., Meagher, C., & Carlson, B. (2008). *The Chicago program evaluation project: A picture of early childhood programs, teachers, and preschool-age children in Chicago*. Princeton: Mathematica Policy Research, Inc. Swank, P. R., Taylor, R. D., Brady, M. P., & Freiberg, H. J. (1989) Sensitivity of classroom observation systems: Measuring teacher effectiveness. *Journal of Experimental Education, 57*, 171–186.

[41] Bryant et al. (2011). Snow, C. E., & Van Hemel, S. B. (Eds.) (2008). *Early childhood assessment: Why, what, and how*. Washington, DC: National Academies Press. Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, OH: Cengage Learning.

[42] Crocker & Algina. (2008). Darling-Hammond, L., Dieckmann, J., Haertel, E., Lotan, R., Newton, X., Philipose, S., Spang, E., Thomas, E., & Williamson, P. (2010). Studying teacher effectiveness: The challenges of developing valid measures. In G. Walford, E. Tucker, and M. Viswanathan (Eds.), *The SAGE Handbook of Measurement* (pp. 87–106). Thousand Oaks, CA: SAGE. Soucaco, E., & Sylva, K. (2010). Developing observation instruments and arriving at inter-rater reliability for a range of context and raters: The early childhood environment rating scales. In G. Walford, E. Tucker, and M. Viswanathan (Eds.), *The SAGE Handbook of Measurement* (pp. 61–86). Thousand Oaks, CA: SAGE. Zieky, M. (2012). *An introduction to evidence centered design for test takers*. Princeton, ETS.

[43] Crocker & Algina. (2008). Darling-Hammond et al. (2010). DeVon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara D. J., Savoy, S. M., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship, 39*, 155–164.

[44] Bryant, D., Maxwell, K., Taylor, K., Poe, M., Peisner-Feinberg, E., & Bernier, K. (2003). *Smart Start and preschool child care quality in NC: Change over time and relation to children's readiness*. Chapel Hill, NC: FPG Child Development Institute. Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-Kindergarten teacher-child interactions and instruction. *Applied Developmental Science, 12*, 140–153. Cassidy, D. J., Hestenes, L. L., Hansen, J. K., Hegde, A., Shim, J., & Hestenes, S. (2005). Revisiting the two faces of child care quality: Structure and process. *Early Education and Development, 16*, 505–520. Cassidy, D. J.; Hestenes, L. L.; Hegde, A.; Hestenes, S.; & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the Early Childhood Environment Rating Scale-Revised. *Early Childhood Research Quarterly, 20*, 345–360. Clifford, Reszka, & Rossbach. (2010). Harms, Clifford, & Cryer. (2005). Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2012). An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development. *Developmental Psychology, 49*, 146–160. Hofer, K. G. (2008). *Measuring quality in pre-kindergarten classrooms: Assessing the Early Childhood Environment Rating Scale* (unpublished dissertation). Nashville: Graduate School of Vanderbilt University. Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*, 732–749. Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., & Yazejian, N. (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72*, 1534–1553. Perlman, M., Zellman, G. L., & Le, V. (2004). Examining the psychometric properties of the Early Childhood Environment Rating Scale-Revised (ECERS-R). *Early Childhood Research Quarterly, 19*, 398–412. Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9*, 144–159. Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Elliot, K. (2002). *Technical Paper 8a: Measuring the impact of pre-school on children's cognitive progress over the pre-school period*. London: Institute of Education, University of London. Zellman, G. L., Perlman, M., Le, V., & Setodji, C. M. (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.

[45] Burchinal et al. (2008). Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly, 25*, 166–176. Curby, T. W., Brock, L. L., & Hamre, B. K. (2013). Teachers' emotional support consistency predicts children's achievement gains and social skills. *Early Education & Development, 24*, 292–309. Curby, T. W., LoCasale-Crouch, J., Konold, T. R., Pianta, R. C., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2009). The relations of observed pre-K classroom quality profiles to children's achievement and social competence. *Early Education & Development, 20*, 346–372. Downer, J. T., Lopez, M. L., Grimm, K. J., Hamagami, A., Pianta, R., C., & Howes, C. (2012). Observations of teacher-child interactions in classrooms serving Latinos and dual language learners: Applicability of the Classroom Assessment Scoring System in diverse settings. *Early Childhood Research Quarterly, 27*, 21–32. Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23*, 27–50. LaParo, K. M., Hamre, B. K., LoCasale-Crouch, J., Pianta, R. C., Bryant, D., Early, D., Clifford, R., Barbarin, O., Howes, C., & Burchinal, M. (2009). Quality in kindergarten classrooms: Observational evidence for the need to increase children's learning opportunities in early education classrooms. *Early Education & Development, 20*, 657–692. Mashburn, A. J., Justice, L. M., Downer, J. T., & Pianta, R. C. (2009). Peer effects on children's language achievement during pre-kindergarten. *Child Development, 80*, 686–702. Mashburn, Pianta, et al. (2008). Pianta, R. C., & Hamre, B. K. (2009). A lot of students and their teachers need support: Using a common framework to observe teacher practices might help. *Educational Researcher, 38*, 546–548. Pianta, Howes, et al. (2005).

[46] Gordon, R., Hofer, K., Wilson, S., & Smith, E. (2012). *Measuring preschool program quality: Multiple aspects of the validity of two widely-used measures (IES proposal).* Chicago: University of Illinois at Chicago. See also http://ies.ed.gov/ncer/projects/grant.asp?ProgID=7&grantid=1451&NameID=351.

[47] Gordon, Fujimoto, et al. (2012).

[48] Hofer, K. G. (2010). How measurement characteristics can affect ECERS-R scores and program funding. *Contemporary Issues in Early Childhood, 11*, 175–191.

[49] Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., Humez, A., Litke, E., & Lynch K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*, 88–106. McClellan, C., Atkinson, M., & Danielson, C. (2012). *Teacher evaluator training & certification: Lessons learned from the Measures of Effective Teaching Project.* San Francisco: Teachscape, Inc.

[50] Archibald, S., Coggshall, J. G., Croft, A., & Goe, L. (2011). *High-quality professional development for all teachers: Effectively allocating resources*. Washington, DC: National Comprehensive Center for Teacher Quality. Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective: Results from a national sample of teachers. *American Educational Research Journal, 38*, 915–945. Jerald, C. D., & Van Hook, K. (2011). *More than measurement: The TAP system's lessons learned for designing better teacher evaluation systems.* Chicago: The Joyce Foundation.

[51] Estacion, A., McMahon, T., Quint, J., Melamud, B., & Stephens, L. (2004). *Conducting classroom observations in First Things First schools*. New York: MDRC.

[52] Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management, 18*, 321–352. Fedor, D. & Rowland, K. (1989). Investigating supervisor attributions of subordinate performance. *Journal of Management, 15*, 405–416.

[53] Harvey, E. A., Fischer, C., Weieneth, J. L., Hurwitz, S. D., & Sayer, A. G. (2013). Predictors of discrepancies between informants' ratings of preschool-aged children's behavior: An examination of ethnicity, child characteristics, and family functioning. *Early Childhood Research Quarterly, 28*, 668–682. Waterman, C., McDermott, P. A., Fantuzzo, J. W., & Gadsden, V. L. (2012). The matter of assessor variance in early childhood education – Or whose score is it anyway? *Early Childhood Research Quarterly, 27*, 46–54.

[54] Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings.* Rockville, MD: Center for Educator Compensation Reform, Westat. Milanowski, A. T., Prince, C. D., & Koppich, J. (2007). *Observations of teachers' classroom performance*. Washington, DC: Center for Educator Compensation Reform, US Department of Education, Office of Elementary and Secondary Education.

[55] Sartain, L., Stoelinga, S. R., & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation.* Chicago: Consortium on Chicago School Research at the University of Chicago Urban Education Institute.

[56] Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill & Melinda Gates Foundation.

[57] Gage, N. L. (2009). *A conception of teaching.* New York: Springer.

[58] Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices (Second edition).* Thousand Oaks, CA: Corwin Press.

[59] Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*, 757–783. Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (in press). *The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol*. Teachers College Record.

[60] Hamre, B. K., Goffin, S. G., & Kraft-Sayre, M. (2009). *Classroom Assessment Scoring System (CLASS) implementation guide*. Charlottesville, VA: Center for Advanced Study of Teaching and Learning.

[61] Heller, J. I., Sheingold, K., & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment, 5*(1): 5–40.

[62] Nijveldt, M., Beijaard, D., Brekelmans, M., Wubbels, T., & Verloop, N. (2009). Assessors' perceptions of their judgment processes: Successful strategies and threats underlying valid assessment of student teachers. *Studies in Educational Evaluation, 35*, 29–36.

[63] Cash et al. (2012). Graham et al. (2012).

[64] Milanowski et al. (2007).

[65] Graham et al. (2012).

[66] Joint Committee on Standards for Educational and Psychological Testing. (1999).

[67] (2003), p. 15.

[68] Brenneman, K., Boller, K., Atkins-Burnett, S., Stipek, D., Forry, N. D., Ertle, B. B., French, L., Ginsburg, H. P., Frede, E., & Schultz, T. (2011). Measuring the quality of early childhood math and science curricula and teaching. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. 77–103). Baltimore: Brookes Publishing.

[69] Bryant. (2010).

[70] Mash, E. J., & McElwee, J. D. (1974). Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. *Child Development, 45*, 367–377.

[71] Graham et al. (2012).

[72] Jerald, C. (2012). *Ensuring accurate feedback from observations: Perspectives on practice*. Seattle: Bill and Melinda Gates Foundation.

[73] http://ersi.info/training_ecers_short_oct2013.html; http://www.teachstone.org/training-programs/regional-training/.

[74] McClellan et al. (2012).

[75] Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*, 371–389.

[76] Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L., & La Paro, K. M. (2007). Classroom assessment scoring system manual, middle/secondary version. Charlottesville: University of Virginia.

[77] Casabianca et al. (2013).

[78] Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., and Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46,* 43–58.

[79] Swenson-Klatt, D., & Tout, K. (2011). Measuring and rating quality: A state perspective on the demands for quality measurement in a policy context. In M. Zaslow, I. Martinez-Beck, K. Tout, & T. Halle, (Eds.), *Quality measurement in early childhood settings* (pp. 347–362). Baltimore: Brookes Publishing.

[80] Wilson, M., & Case, H. (1997). An examination of variation in rater severity over time: A study in rater drift. Berkeley, CA: Berkeley Evaluation and Assessment Research (BEAR) Center, University of California, Berkeley.

[81] Hamre, B. K., Goffin, S. G., & Kraft-Sayre, M. (2009). *Classroom Assessment Scoring System (CLASS) implementation guide: Measuring and improving classroom interactions in early childhood settings*. Charlottesville, VA: CASTL.

[82] Joe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). *Foundations of observation: Considerations for developing a classroom observation system that helps districts achieve consistent and accurate scores*. Seattle, WA: Bill & Melinda Gates Foundation.

[83] Caronongan, P., Kirby, G., Malone, L., & Boller, K. (2011). *Defining and measuring quality: An in-depth study of five child care quality rating and improvement systems* (OPRE Report 2011-29). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

[84] For a discussion of how many classrooms can serve as an adequate sample, see Hamre & Maxwell. (2011). Kemper, E. A., Stringfield, S., & Teddlie, C. (2003). Mixed methods sampling strategies in social science research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 273–296). Thousand Oaks, CA: Sage Publications, Inc.

[85] Karoly, L. A., Zellman, G. L., & Perlman, M. (2013). Understanding variation in classroom quality within early childhood centers: Evidence from Colorado's quality rating and improvement system. *Early Childhood Research Quarterly, 28*, 645–657.

[86] Hill, Charalambous, & Kraft. (2012). Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum in large-scale surveys: A study of literacy teaching in third-grade classrooms. *Elementary School Journal, 105*, 75–102.

[87] Bell, C. A., Gitomer, D. H., McCaffrey, D., Hamre, B., Pianta, R., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87.

[88] Ho & Kane. (2013).

[89] Meyer, J. P., Cash, A. H. & Mashburn, A. J. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment*, *16*, 227–243.

[90] Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment, 13*, 267–300.

[91] Hill, Charalambous, Blazar, et al.

[92] Hill, Charalambous, & Kraft. (2012).

[93] Hill, Charalambous, & Kraft. (2012). Hintze. (2005).

[94] Stodolsky, S. S. (1984). Teacher evaluation: The limits of looking. *Educational Researcher*, *13*(9), 11–18.

[95] Rowan et al. (2004).

[96] Brenneman et al. (2011). Sylva, K., Siraj-Blatchford, I., & Taggart, B. (2011). *ECERS-E: The four curricular subscales extension to the Early Childhood Environment Rating Scale (ECERS-R) (4th Edition)*. New York: Teachers College Press.

[97] Booren, L. M., Downer, J. T., & Vitiello, V. E. (2012). Observations of children's interactions with teachers, peers, and tasks across preschool classroom activity settings. *Early Education and Development, 23*, 517–538. Buell, M. J., May, H., Han, M., & Vukelich, C. (2013). *Exploring variance in the Pre-K Classroom Assessment Scoring System (CLASS) across classroom context*. Presentation at the Annual Meeting of the American Educational Research Association, San Francisco. Cabell, S. Q., DeCoster, J., LoCasale-Crouch, J., Hamre, B. K., & Pianta, R. C. (2013). Variation in the effectiveness of instructional interactions across preschool classroom settings and learning activities. *Early Childhood Research Quarterly,* 28, 820–830. Vitiello, V. E., Booren, L. M., Downer, J. T, & Williford, A. P. (2012). Variation in children's classroom engagement throughout a day in preschool: Relations to classroom and child factors. *Early Childhood Research Quarterly, 27*, 210–220.

[98] Curby, T. W., Grimm, K. J., & Pianta, R. C. (2010). Stability and change in early childhood classroom interactions during the first two hours of a day. *Early Childhood Research Quarterly, 25*, 373–384.

[99] Curby, T. W., Stuhlman, M., Grimm, K., Mashburn, A., Chomat-Monney, L., Downer, J., Hamre, B., & Pianta, R. C. (2011). Within-day variability in the quality of classroom interactions during Third and Fifth Grade. *The Elementary School Journal, 112*(1), 16–37.

[100] Hamre & Maxwell. (2011).

[101] Barnett et al. (2011).

[102] Smith, M. W., Brady, J. P., & Anastasopoulos, L. (2008). *Early Language and Literacy Classroom Observation Tool, Pre-K (ELLCO Pre-K)*. Baltimore, MD: Brookes Publishing.

[103] High/Scope Educational Research Foundation. *Preschool Program Quality Assessment (PQA)*. Ypsilanti, MI: High/Scope Press.

[104] Layton, L. (2013). Education Secretary Arne Duncan works to sell Obama administration's preschool initiative. *Washington Post (online edition)*, retrieved June 13, 2013 from http://www.washingtonpost.com/politics/education-secretary-arne-duncan-works-to-sell-obama-administrations-preschool-initiative/2013/06/12/ba25e6a4-cd2e-11e2-8845-d970ccb04497_story.html.

[105] Includes professional development, technical assistance, and/or equipment and materials.

[106] Includes ensuring scores meet criteria for state QRIS or NAEYC accreditation.

[107] Districts may choose from among these protocols.

[108] ECERS-R must be utilized at least once every three years; additional protocols may be used in other years.

About ETS

At ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research. ETS serves individuals, educational institutions and government agencies by providing customized solutions for teacher certification, English language learning, and elementary, secondary and post-secondary education, as well as conducting education research, analysis and policy studies. Founded as a nonprofit in 1947, ETS develops, administers and scores more than 50 million tests annually — including the *TOEFL*® and *TOEIC*® tests, the *GRE*® tests and *The Praxis Series*™ assessments — in more than 180 countries, at over 9,000 locations worldwide.

22802

ETS®

Listening. Learning. Leading.®

www.ets.org