USES OF DATA ON CHILD OUTCOMES AND PROGRAM PROCESSES IN EARLY CHILDHOOD ACCOUNTABILITY SYSTEMS: ASSUMPTIONS, CHALLENGES, AND CONSEQUENCES

John M. Love
Mathematica Policy Research, Inc.
Princeton, New Jersey

September 25, 2006

Prepared for the National Task Force on Early Childhood Accountability, with support from The Pew Charitable Trusts, the Foundation for Child Development, and the Joyce Foundation. I am greatly indebted to Tom Schultz of the Pew Charitable Trusts for challenging my own assumptions, taking the time to discuss this topic, and for his many helpful suggestions on earlier drafts. Responsibility for any shortcomings remains with me, however.

Abstract

Establishing state early childhood accountability systems requires careful consideration of who will be using the results of the assessments and how the results are likely to be used. Many benefits can accrue to well-conceived accountability systems, and in considering the issues of uses and users of the data, this paper assumes that (1) the key users will be at the state level, (2) accountability systems will be standards based, and that (3) there are multiple uses and users. A number of difficult challenges confront those who would establish a successful accountability system. This paper focuses on challenges associated with attributing measured outcomes to characteristics of the programs, implementing a large system with limited resources, and anticipating the potential negative unintended consequences; it also notes other challenges. The decisions states make about their system will dramatically affect their ability to achieve their goals; three of the key decisions are highlighted: (1) level of data collected (program and/or child), (2) nature of the outcome (status or change over time), and (3) sampling (everyone or a sample). Several actions are described that can safeguard the integrity of the accountability system. After providing some concrete examples of what a state accountability system might look like under different scenarios of users and their main uses, the paper concludes by drawing all these considerations together into five sets of actions that can lead to an effective accountability system. The result will be a system that meets the needs of stakeholders and has the greatest potential for creating programs in which all children will be entering school on trajectories for success.

Introduction

Accountability is not a new topic for the early childhood field. It is taking new and more urgent forms today, but it may be useful to recall that the enactment of Head Start 41 years ago "coincided with another development: a new demand for 'accountability' for demonstrable 'outputs,' and measurable results of government spending" (Rivlin 1974, p. 10). In the 1960s, Robert McNamara's planning, programming, and budgeting systems installed at the Pentagon influenced most federal agencies. Soon, Congress and the Bureau of the Budget demanded accountability (Datta 1976). More recently, the *Government Performance and Results Act of 1994* (GPRA) required all federal agencies to develop performance measures to chart the performance of their programs. As this country's major federal early childhood program, Head Start responded with a comprehensive child outcomes framework to guide performance assessments (Administration on Children, Youth and Families 2000). As we all know, approaches to accountability within Head Start today continue to inspire debate. In the K-12 education system, the *No Child Left Behind Act* holds states accountable for the academic progress of all students.

As states make decisions about the shape of their early childhood accountability systems, many aspects of such systems come under scrutiny, as seen in the goals and deliberations of the National Task Force on Early Childhood Accountability. Typically, state system elements include (1) the expected child outcomes, often operationalized as "early learning standards," to which accountability systems should be aligned; (2) the standards set for program operations and quality; and (3) the measures that are used for assessing children's early development and learning and the quality (process data) of the programs. As decisions about these elements are

being considered, it is crucial to think through the implications of the intended uses of the information that emerge from any accountability system. My hope is that the assumptions and strategies for action laid out here will be useful to all stakeholders in state prekindergarten programs who want to create and maintain ways of ensuring high quality programs for children in the year or two before kindergarten entry.

My first thoughts when the subject of accountability comes up are about the consequences. What distinguishes an accountability system that encompasses program and child assessments is that the results of the assessments are interpreted in ways that allow the responsible agency or agencies to make decisions that will change something about the programs. We generally assume that the intention of these changes is to make programs better. My hope that these decisions will lead to better programs is grounded in the belief that, for the most part, the accountability movement is fueled by a genuine desire to improve early childhood services so that all children have a better chance of realizing their potential to succeed in school.

Potential benefits of accountability systems. Before the positive outcomes from establishing accountability systems can be realized, many issues need to be addressed and important challenges anticipated and overcome. But let's hold off considering the challenges and look first at the wide range of benefits that have been claimed for accountability assessment systems. Do the following potential benefits suggest that the challenges and struggles will be worth the effort?

- ***Program improvement.*** This is perhaps the most important benefit, the outcome that is touted as a major rationale for having accountability systems. Data about the areas in which children show gains from their program participation can guide programs in disseminating the program features that are associated with gains. When results show that a failure to improve child outcomes is associated with particular program characteristics, technical assistance goes into full swing to address the programs' shortcomings.
- ***Positive curriculum changes.*** One of the most important changes that could come about would be in the area of curriculum or instructional approaches. I'm sure that the

astute, entrepreneurial curriculum developers will be closely watching state accountability efforts for signs that their particular curriculum will contain the seeds to success. State program managers will undoubtedly have a rich menu of options from which to choose more-effective curricula.

- ***Enhanced professional development.*** Evidence of program areas needing strengthening can lead program leaders to address problems with targeted improvements. A critical element of many program improvement efforts, developing the early childhood workforce, has particular salience because of the potential ripple effects by which staff enhancements benefit the larger network of state programs.

- ***More-effective resource allocation.*** Program quality is often associated with how state resources are allocated (to training and technical assistance, monitoring and oversight, inspections or support). Properly analyzed accountability data can better inform state agencies about where resource allocations are most needed and can have the largest payoff.

- ***Reduction in bureaucratic rules.*** Schorr, Farrow, Hornbeck, and Watson (1994) argued that a major benefit of results-based accountability is that because of new data available about what really matters—the quality of the programs and the outcomes for their children—states can eliminate, or at least reduce, their need to enforce rigid rules about service delivery: "Management by results is the best alternative to top-down, centralized micromanagement, which holds programs responsible for adhering to rules…that interfere [with their] ability to respond to a wide range of urgent needs" (p. 2).

- ***Monitoring trends over time.*** Systematic data on program characteristics, processes, and child outcome measures allow states to examine in detail the operations and outcomes of their programs. For example, Denno, Miller, and Joyce (2005) describe how the Ohio early learning accountability system allows the state department of education to monitor programs' progress across years.

- ***Enhanced support for early childhood programs.*** It is plausible to assume that the credibility of early childhood programs will be enhanced if state policy leaders see the programs as being held accountable for results—as well as demonstrating positive outcomes for children over time. In their review of universal prekindergarten efforts, Christina and Nicholson-Goodman (2005) found states in which the increased emphasis on accountability has heightened awareness of how important these programs are.

Enhancement of children's learning and development. The bottom-line intended consequence is that all children served by the programs will be better off. This means they will develop more fully along all the dimensions the program seeks to enhance, and they will be launched on learning trajectories enabling them to succeed in their next school experience.

I will revisit some of these beneficial consequences of accountability systems in the context of discussing who the users are.

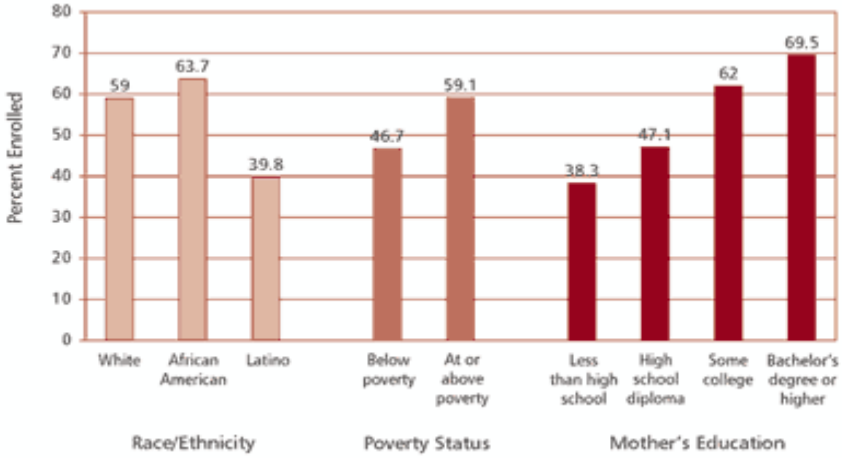
Breadth of impact of accountability decisions. Accountability system decisions have the potential to affect every state in the union, tens of thousands of programs, and millions of children. State funding for prekindergarten has grown in the past 10 to 15 years, and continues growing, even while some sources of federal support diminish. Fifteen years ago, only 24 states funded prekindergarten classrooms, but by 1998-1999, 42 did (Rosenthal, Rathbun, & West 2005). But still, not all children are enrolled, even in states with such programs, nor do the programs provide extensive services. The Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) found that of all children in the kindergarten class of fall 1998, 68 percent had participated in preschool in the previous year (i.e., in center-based early care and education programs or Head Start) (Rosenthal et al. 2005). But children participated on average only 22 hours per week. The National Center for Education Statistics' analysis of National Household Education Survey reports somewhat lower numbers: in 2005, 55 percent of 3- to 5-year-olds were in a center-based program (Iruka & Carver 2006).[1]

Also relevant to the implementation of the task force recommendations are the substantial variations in this pattern found across the country. The percentage of children attending preschool ranged from 62 percent in the Western states to 72 percent in the Midwest. Mean hours per week were highest in the South (28 hours per week) and lowest in the West (20 hours).

Preschool enrollment rates also vary among children who differ in terms of race/ethnicity, poverty status, and family income, with higher enrollment rates for African American children, children in higher-income families, and children whose mothers have higher education levels:

[1]The latest report from the National Center for Education Statistics' analysis of the 2005 National Household Education Survey's Early Childhood Program Participation Survey includes data from parents or guardians of 7,198 children under age 6 (which when weighted represent 20.7 million children under 6 not enrolled in kindergarten). Center arrangements include cay care centers, Head Start and Early Head Start programs, preschool, prekindergartens, and other early childhood programs. The percentage of children in *all nonparental* care

Figure 1: Preschool program enrollment by child and mother characteristics[2]



Obtaining more-recent data will be useful as the task force considers how its recommendations might be received—and acted on—in various regions.

Assumptions Underlying Accountability Systems

Before the hoped-for benefits can be realized, we should think about the assumptions underlying effective implementation of accountability systems. We can begin with the fact that accountability requires information and that the information most likely to be of value to those responsible for the effectiveness of state prekindergarten programs comes from assessments—assessments of children and assessments of programs. In the view of the National Education Goals Panel, which is responsible for today's widely held conception of school readiness, child assessment has five purposes, one of which is "high-stakes accountability" (Shepard, Kagan, & Wurtz 1998). But accountability can mean different things to different users of the information. To one, it could mean collecting child-level data for deciding the fate and trajectories of

(continued)
arrangements, including relative and nonrelative home care, was considerably higher: 73 percent of all children between the ages of 3 and 5 (Iruka & Carver 2006).

individual children; for others, it means obtaining program-level data for deciding whether a program should continue operation, receive remedial attention, or be discontinued. Because the user of the assessment information is so central to any discussions about how the data might be used, I begin by stating my assumptions about who the users are. Following that, I describe additional assumptions that often are seen in accountability systems, and that are important for the task force to consider.

States will be the most important users of accountability information. Although early childhood programs operate through many funding streams, it is the states that are, or will be, the central focus for shaping policies related to accountability. With the exception of Head Start, all major federal early childhood programs give responsibility to state agencies to set standards, monitor local programs, and make funding decisions. Thus, states are also the public entities most likely to take the leadership role in designing and implementing early childhood accountability systems. Many states already have established early learning standards for their prekindergarten programs (Scott-Little, Kagan, & Clifford 2003) as well as minimum standards for program quality and have the greatest stake in implementing assessments to hold programs accountable for meeting those standards.

There is great diversity across states in the populations they serve, how they define their early learning benchmarks, the political forces they need to respond to, and so forth. If learning standards differ, accountability assessments should differ. Thus, many features of an accountability system must be tailored to local circumstances. If learning standards for the elementary grades differ across states, then what constitutes a "ready" child may differ, and child assessments would likewise differ across states.

(continued)

[2]From *Harvard Education Letter*, July/August 2005, which was adapted from Karoly and Bigelow (2005).

Another reason accountability systems need to be tailored to their locale of implementation is that the program (or state) response to the assessment results must take into account characteristics of the prekindergarten systems. As Gilliam and Marchesseault (2005) show, tremendous variation in teacher qualifications exists across state programs. For example, although, on average, 49 percent of teachers report having a bachelor's degree, the majority of teachers in five state systems hold master's degrees and in three systems the majority have only a high school diploma or GED (out of the 52 state-funded prekindergarten systems that Gilliam and Marchesseault sampled). The extent of inservice training also varies widely, to give another example: annual prekindergarten system hours of inservice training (as reported by the teachers interviewed) range from a low of 23 hours to a high more than double that, 55 hours. The systems' investments in inservice training may suggest the availability of resources for acting on accountability results. If a newly developed accountability system produces findings that must be acted on, knowing something about the skills—and availability—of the actors is important.

Absent a review of 50 state early learning standards telling us that there is a high degree of congruence, I assume some differentiation. This has implications for the accountability assessment enterprise. Systems need to be tailored to the characteristics of each state. The important characteristics include their prekindergarten learning standards, their goals for accountability, purposes to which they might put the child and program assessment data, and so forth. Not everyone agrees with this assumption. Some writers are questioning the existence of 50 different standards. For example, with respect to NCLB, Diane Ravitch (2005) recently has argued that it does not make any sense for our country to have different standards across states. For her, we should have the same goals for all U.S. children, whether they happen to live in California, New Jersey, North Dakota, or the District of Columbia. If a move toward national standards ever came about, we would see serious implications for accountability systems. This

would not, however, alter the fact that attention to the needs of state decision makers is a central assumption for effective accountability.

Assessment decisions are linked to clear expectations for children's early learning and development. Before an assessment system is designed, whether for accountability or any other purpose, important stakeholders should agree on what is to be assessed. If this were a program evaluation discussion, we would move immediately to creating a logic model to show the program's theory of change, that is, what it expects to achieve and by what means. That notion is useful for accountability discussions as well. But the main point is to have a target with which the assessments can be aligned. Early learning standards, which generally refer to expectations for children's learning and development (National Association for the Education of Young Children [NAEYC] 2002), provide such a target. Sometimes these are called content standards, benchmarks, or performance standards. By any name, they delineate what preschool children are expected to know and be able to do by the time they complete their prekindergarten years. As of 2002, three-fourths of states had adopted or were developing such standards (Scott-Little, Kagan, & Frelow 2003). Without a clear sense of what outcomes the state values, the results of child assessments will not be as useful to the decision makers. As the National Association for the Education of Young Children (NAEYC) puts it, "mismatches between program goals and evaluation design and instruments may lead to erroneous conclusions about the effectiveness of particular interventions" (NAEYC 2003, p. 14).

Prekindergarten standards do not exist in isolation. States typically align their early learning standards with already-developed K-12 standards. The National Governors' Association (NGA), in fact, strongly encourages such alignments, assuming that "when aligned with expectations for kindergarten entry, early learning standards can help ensure that children start school ready to achieve their full potential" (NGA 2003, p. 2).

States also may create *program* standards, which refer to the expectations for the quality and operations of the programs. To the extent that accountability systems incorporate assessments of program features (whether global quality, specific teacher-child interactions, or simply structural characteristics such as ratios and group sizes), knowing the state's standards for programs also is essential.

Knowing the u*sers* of the information is essential for thinking about us*es* of the information. The third assumption recognizes that different users of assessment information are concerned about different key questions, have different goals, will use information to guide different types of decisions, respond to pressures from different constituents, and may have different philosophies about both early childhood programs and accountability. Users also differ in their technical sophistication—the extent to which they understand assessment, know measures, and understand appropriate and inappropriate interpretations and uses. Table 1 suggests some possible scenarios of intended users of assessment data and how they might use the information..

I want to emphasize four points about these scenarios: First, I included the local programs and their communities below the dashed line. These are of course very important stakeholders in the overall accountability process. Many accountability strategies presume that documenting and publicizing program outcomes will create powerful incentives for locally driven program improvement. But for the purposes of this paper, I consider local staff to be secondary users. Their needs and interests should be taken into account at some level, but our first priority is to consider the state-level decision makers. Second, there is considerable overlap across the users, both in their questions and in the potential uses they have for data. If this reflects reality, it means that we do not need to take into account as many contingencies in considering how the remaining

Table 1. Potential users of accountability systems, their main questions, and potential uses of the data

Users	Questions	Potential Uses
State oversight and funding bodies, e.g., legislatures and other high-level decision makers	How are all young children doing in terms of trends in school readiness and success in kg.-grade 3 schooling? Does the state prekindergarten program work? Are state programs as a whole producing the outcomes we want for our children? Does it work well enough? Are the programs good quality? Who is responsible for the poorly performing programs? Do programs of greater intensity or duration produce more-positive outcomes (full-day vs. part-day, one vs. two years of prekindergarten)?	Track use of funds Justify budgets Make new-budget decisions Increase funding Decrease or stop funding Fund alternative programs Modify/refine state early childhood policies Modify prekindergarten curricula to create better prekindergarten-K-3 articulation
State-level program managers	How are programs doing, overall? Which programs require attention (additional resources, T/TA support, facilities improvements)? Are there groups of children who require more attention? Are some types of programs more successful in promoting success for low-income, minority, and English learners, or children with disabilities?	Convince legislature that pre-k is "working" Contribute to state planning and system design Monitor local programs Defund programs Set priorities for program improvement Provide feedback to local program managers Allocate program improvement (e.g., T/TA) resources Create (or modify an existing) incentives system Guide professional development Conduct public awareness campaigns; work with the media Revise/modify/improve the accountability system and its components (child assessments, program quality rating systems, etc.)
Local program managers; local communities	How do my program's quality and outcomes compare with other programs statewide? How are my teachers doing? Which teachers need help? Is the local curriculum accomplishing its goals? Which teachers could serve as models for staff development for others? Do certain groups of children need more help?	Strengthen or modify inservice training programs Consider other means for program improvement Provide incentives to certain teachers Modify/revise program curriculum Modify instructional approaches overall Modify instructional approaches for particular groups of children

assumptions play out for the various user groups. Third, the questions of interest for different users have significant implications for the scale and types of required assessment data. For example, assessments from a random sample of prekindergarten programs within the state may be sufficient to answer questions from state legislators on the typical outcomes of the program. On the other hand, if a state program manager wants to know whether any local programs require additional help, data must be collected for each local program.

Finally, none of the uses outlined in Table 1 requires individual child-level data. This is in itself an important assumption, which has implications for system design. As can be seen, a wide range of important decisions can be made without having data reported at the individual child level. But one use that becomes impossible under this scenario is making decisions about individual children, such as decisions about their placement or what instructional approaches to take.[3] Nevertheless, this does not (at least in my judgment) mean that such a system is not high stakes. As the uses listed in the right-hand column show, crucial decisions with high stakes for *programs* and their managers and staff can also be made on the basis of group data (classroom-, center-, or program-level information).[4] For example, Deborah Stipek (2005) predicts that, given language currently in the Congressional bills for Head Start reauthorization, Head Start programs could be held accountable for making progress toward new education performance goals, with funding eventually withdrawn if they fail to show progress.

Issues of knowing the users become more complex as we realize that state accountability systems will have multiple users and stakeholders. As Table 1 suggests, the same results may be

[3]This does not mean that individual child-level data and decisions are impossible in a statewide system (see, for example, Grafwallner, 2005), but such uses introduce another entire set of issues that require separate attention.

[4]Some writers have defined high-stakes accountability as that in which assessments are used to make decisions about individual children or teachers (e.g., Maxwell & Clifford 2004). I believe other decisions can have high stakes as well.

acted on by users with vastly different agendas, from legislative initiatives regarding credentialing to program- and classroom-level improvement efforts. This increases the chances for unintended consequences.

Challenges to Creating Effective Accountability Systems

I want to discuss three major challenges—attributing the results to the programs, coping with limited resources for implementing accountability, and avoiding unintended consequences—and then list a number of other challenges to be aware of. These challenges exist across all the types of users and forms of uses outlined in Table 1.

Interpretation of assessment data: the attribution question. Results of the assessments need interpretation before they can be used by at any level—by local programs, state legislatures, or state program managers. Users want to hold programs accountable for their results, but the results of the assessments alone do not allow us to conclude that it is the *programs* that caused the observed results. Even when states focus on children's progress (change over time), the question remains: what (or who) is responsible for that progress?

Although an accountability assessment system is not a program evaluation, certain parallels are inescapable. Examining these parallels helps to highlight the issue. Both begin by defining expected outcomes and have hypotheses about expected results. Both compare the outcomes for the group of interest against some benchmark. In an evaluation, the benchmark is typically a control or comparison group—children intended to be just like those the evaluators are interested in except that they do not receive the intervention. In a statewide accountability system, there is no similar contrast group because all (or potentially all) eligible children in the state are enrolled in the program. However, the state decision makers assume that the outcomes can be attributed to the programs the children are in. Furthermore, when state agencies want to know about the relative success of different programs, or want to hold programs accountable for the progress of

different groups of children, the situation parallels one in which a program evaluation is designed to learn what works for whom under what conditions.

A program evaluation, however, would ensure that these comparisons are fair and unbiased by implementing such strategies as randomly assigning children to different programs or at least controlling for relevant background factors when testing for differences between groups. The evaluator would think about how to ensure that the gains of English learners, to take just one of the policy relevant subgroups of children, were contrasted with gains for an appropriate comparison group of English learners. Thus, in important ways, decisions that states may be expected to make based on their accountability assessment results run the risk of having to be made with insufficient evidence.

This is a serious dilemma, but a solution may lie in multidimensional approaches. States could, for example, use their accountability system for obtaining data across all elements of their prekindergarten program (to the extent needed to answer the relevant questions, such as those in Table 1) and also conduct small-scale, focused experimental studies that could provide answers to particular state policy questions. For example, a question about program intensity or duration would better be addressed through experiments that could systematically vary the length of the program day for relatively small samples of children to learn whether that makes a difference in children's kindergarten readiness outcomes. Statewide accountability data could be used to get preliminary answers to these questions, while smaller-scale experiments could provide stronger evidence for understanding the nature of the patterns.

A second way in which accountability systems can achieve their purposes entails focusing the questions on absolute standards of progress, or criterion-referenced data, rather than questions about relative progress that require comparisons across different subgroups or segments of the prekindergarten system or population. For example, such questions in Table 1

as “How are all young children doing in terms of trends in school readiness?” and “Are state programs producing the outcomes we want for our children?” can be answered by reference to absolute standards. The standards could be the norms of a particular assessment instrument (are the state’s 5-year-olds, on average, functioning at the test’s 5-year-old level?). Or, children’s performance could be compared with the state’s early learning benchmarks or standards (do the children understand that reading progresses from left to right and from top to bottom on the pages?). These analyses could provide meaningful descriptions of children’s developmental progress, yet would not be able to attribute the progress to the program.

In contrasting three accountability models, the Council of Chief State School Officers (CCSSO) provides a useful synopsis of the options. Goldschmidt et al. (2005) describe status, improvement, and growth models, with the value-added model (VAM) presented as one type of growth model. Using VAM, a state (or smaller unit) uses statistical controls to isolate the effects of particular programs on children’s progress. “The main purpose of VAMs is to separate the effects of non-school-related factors (such as family, peer, and individual influence) from a school’s performance at any point in time *so that student performance can be attributed appropriately*” (Goldschmidt et al., 2005, p. 5, emphasis added).

As I will discuss later under implementation strategies, what is missing from this, and most accountability formulations, is the crucial role of process data. If the state obtains data on program processes (the context of classrooms, teachers, and families) so that outcomes can be examined *in the context of how well programs are implemented*, this strengthens the case for holding programs accountable for their outcomes.

Thus, (1) when accountability data are criterion referenced; (2) are based on children’s progress over time; (3) are subjected to statistical controls for the characteristics of the children

and families that may be associated with the kinds of gains children make while in the program;[5] and (4) can be related to evidence about the nature of program implementation, then an accountability system can be used for holding programs responsible for children's progess. Then if the appropriate analyses are done, the state has a strong basis for concluding that children's progress is more strongly associated with the programs they attended than with the various characteristics the children brought to the programs in the first place.

Implementing high-quality data collection in the face of limited resources. The typical program evaluation has a very high per-child cost for its design, data collection, and analysis. It carefully considers the universe that it is charged with evaluating and finds ways to select a sample that will represent the universe that policymakers want to learn about, randomly assigns participants to program and control groups, and often uses child and classroom measures and data collection methods that could not be afforded if every classroom and child in the state were being assessed. Although states may want to come close to the quality system that an in-depth evaluation can employ, compromises will be necessary. Child assessments may need to be administered by teachers, so have to be less complex, require less-rigorous and time-consuming training, and be designed to leave less room for interpretation. State personnel may conduct the classroom observations, so the observation system needs to be one that can be learned in a short time, with rating categories as objective and low-inference as possible. Finally, as in most research, sampling (of program units and children within local programs) can be judicially used to reduce costs.

[5]Another challenge for states the lack of sufficient data for introducing the statistical controls. Analyses can control only for characteristics that have been measured, and it is rare that program data systems contain all the relevant information on phenomena that may be influencing children's progress in the programs.

Avoiding unintended, negative consequences. We hope negative consequences will not occur, but we all know there is a chance they will crop up with an unknown probability, even with the best of intentions. Identifying potential negative consequences in advance will help the system developers, program managers, and decision makers guard against them. I list some of the more likely—or at least feared—unintended consequences:

- ***Misusing child outcome data.*** If data are not interpreted carefully, it is possible that program managers, at either the local or state level, might penalize programs that are serving the most vulnerable children, including those at the highest levels of risk and English learners for whom the assessments are not so valid (see NAEYC 2003, p. 14). This could inadvertently create an incentive for programs to recruit only more-advanced children.

- ***Less-challenging standards.*** Schorr et al. (1994) warn that one risk is that funders (states) could "confine their support to interventions whose effects are readily and quickly quantifiable" (p. 6). Standards could be modified either by writing benchmarks that are easier to achieve or by setting the cut-points lower.

- ***Easier or less-comprehensive assessments.*** This is an especially likely consequence when so many choices for child assessments abound and it is possible for states to adopt those that are inexpensive, quick, and yield simple numbers.

- ***Curriculum changes.*** Just as an improved curriculum is a hoped-for benefit, a worse curriculum is a risk. There may be some temptation to implement teaching approaches that conform to assessment items, as in "teaching to the test" (see Meisels, Atkins-Burnett, Xue, Bickel, & Son 2003).

- ***Shifting resources from the program's main purposes.*** The time-consuming requirements of a high-profile accountability system (including administering, recording, and reporting mandatory assessments of children, classrooms, and other program attributes) could lead local programs to shift time and attention of staff and managers away from efforts to further improve teaching and learning. The potential positive impact of using the results of accountability systems to improve programs may not be realized if there is a reduction in available staff time and leadership.

Defunding programs with potential. If assessment results are examined only superficially, some programs may look like they are not effective when in fact they are working with difficult-to-teach children who face multiple risks, or face other extenuating circumstances. How do we keep states from "shooting from the hip" as soon as they see data that make a program look bad?

Increased awareness of these potential negative consequences is a first step in avoiding them. This in itself is a major challenge. Several additional challenges need to be considered as well.

Building awareness and knowledge within the state. This is important for obtaining widespread acceptance among all early childhood stakeholders so that a climate of acceptance exists when results—and consequences—emerge. Awareness of the system details and knowledge of how it works can prepare decision makers for the crucial task of recommending (or taking) actions based on the assessment results.

Creating an infrastructure to support the system. There are several pieces to infrastructure development, but two important ones are creating a large-scale data collection system that ensures complete and reliable data and identifying (or developing) staff with the expertise to conduct the necessary analyses and report their findings clearly to agency decision makers. Ensuring the necessary financial support is another aspect of the infrastructure.

Making wise decisions about child and program measures and assessment procedures. It is not a trivial task to identify the best measures (whether child development or program process measures) from the very large number of imperfect assessment instruments currently available. In addition, as noted, it is challenging to go through the process needed to ensure that the measures align with the state early learning standards or their expectations for children's learning and development.

Preparing for use. As soon as that first report comes in from the statewide data collection, someone needs to act. States should begin by gathering the expertise needed for implementing actions in response to assessment results. They should be prepared to genuinely address the shortcomings identified, not over-react, and avoid the negative unintended consequences. One of the more important potential uses, which takes considerable preparation, is modifying

programs that appear not to be performing adequately. Having access to well-developed (and tested) curricula and a technical assistance network that can ensure their implementation should be high on the "preparation" agenda.

Recommended Implementation Strategies

In some sense, all of the forgoing is necessary for implementation. But I see four aspects of creating an accountability assessment system that involve clear choices: (1) obtaining child or program data or both; (2) measuring children's end-of-program status or gains in the program; (3) analyzing outcomes only or in the context of process; and (4) assessing all children and programs or sampling. These four strategies apply across the full set of uses and users. Although different users will design their accountability or assessment strategy differently, these four are appropriate for all users and uses. Which direction states take can profoundly affect both the implementation and use of data from their accountability system. As I discuss these four system aspects, it is clear what I recommend; however, each state must weigh these and make its own choices.

Program and child data: the solution is not either-or. There is a long history of systems relying on process data to hold programs accountable for delivering needed services. As Lisbeth Schorr and others have so eloquently argued, an outcomes-based approach to accountability is essential for guiding changes if they are to have some chance of improving the outcomes. In fact, Schorr et al. argue that "results-based accountability is an *essential* part of a larger strategy to improve outcomes for children" (Schorr et al. 1994, p. 2, emphasis added), largely because, in their view, the focus for programs shifts from "'Did you do what they told you to do?' to 'Did it work? What difference did it make in outcomes for children?'" (Schorr et al. 1994, p. 3). We can improve only what we know about; knowing that a program provides *x* amount of space per child allows us to change the amount of space in some desired direction if *x* is not consistent with

the program's expectation. However, a focus on process (or in some cases, "inputs") tells decision makers nothing about whether a program is achieving its goals. Today it is rare to hear discussions of accountability that do *not* include acknowledgment of the need for data on the outcomes that the programs are responsible for achieving. Funding decisions, for example, can be made more wisely if both program process and child outcome data are considered.

Measure both children's status at the end of the program and their growth in the program. A variety of methods can be used to characterize a program's success in meeting the state's objectives (Linn, Baker, & Betebenner 2002). Four of the main ways results could be reported include (1) use of a cut score to report the percentage scoring above a given level; (2) use of multiple cut scores so that results could be in terms of the percentage scoring above, for example, a basic level, minimal level, proficient level, and so on; (3) index scores—Linn et al. (2002) describe these as based on a metric essentially assigning partial scores to children who score at some level below "proficient," for example, those in the proficient range get 1.0, those at the high end of the basic range get 0.8, those at mid-range get 0.6, and so forth; and (4) reporting gains using a metric, such as standard-deviation units (effect sizes), that provide meaningful information that can be compared across measures and groups. In one state, the state budget office may be interested only in the percentage of children scoring above a cut score established to reflect the standard of "readiness" desired for kindergarten entry. In another state, the user at the state level may be more interested in how much children have learned during their prekindergarten year so that budget resources can be allocated to help those programs where children are not showing gains, for whatever reason.

Programs that produce gains in children's performance are to be encouraged. Nevertheless, we must realize that programs whose children enter kindergarten performing at a high level *may or may not* be doing a good job; it depends on where the children started. The same status may

be seen in an excellent program that created large gains in low-performing children and in an ineffective program that happens to enroll a lot of children who were performing well at the outset. Gain scores are necessary to tell the difference. On the other hand, status at the end of prekindergarten—or, more typically, at kindergarten entry—is an important goal in itself, so states might want to consider status measures as well. Even programs working with the most difficult-to-teach children need to be encouraged to do even better when it is found that the gains they achieve are not sufficient to move children onto a trajectory for success in school.[6]

The same issue exists with respect to program data. Programs need to improve, but it is not enough to be satisfied with improvements that fall short of achieving the level of quality programming that is necessary to meet children's needs. Thus, we can conceive of applying some absolute standards for program features and not be satisfied with programs simply getting "better." A state early childhood program manager may want to create a media campaign to promote the levels of process quality once they have been achieved in some of the state prekindergarten programs.

Analyze assessment data in ways that permit true accountability. Programs cannot be held accountable for their performance if it is unclear whether their performance is responsible for the outcomes. Unfortunately, accepting the importance of building an accountability system on program outcomes has sometimes led to an over-reaction: that is, there is a risk of relying on outcome data to the exclusion of process or input data. *Both* are needed. States need to be able to hold programs accountable both for implementing the kinds of programs that the designers believe will lead to the intended outcomes and for achieving those outcomes. To accomplish this,

[6]This introduces some complex issues (beyond the scope of this paper) relating to who is responsible for children's ultimate success in school. This paper focuses on ways of using data to help ensure responsible prekindergarten programs. However, those who are responsible for the children after they enter kindergarten may need to increase their own effectiveness to maintain the trajectories that good preschool programs begin.

decision makers must have the ability to interpret the program's outcomes *in light of* program process data. Some might argue that since achieving the state learning standards (or benchmarks, or expected outcomes) is what is important, it doesn't matter how the programs do it. But another concern of the task force, along with many other writers on accountability issues, is being able to use the results for program improvement. This requires the ability to link outcomes with process, to know analytically whether (and how) the observed changes in children's performance are related to the experiences they had in the program. One characteristic of the Head Start NRS, for example, is that the Head Start Bureau receives child outcome data summaries in isolation from knowledge about program characteristics other than demographics (Government Accountability Office 2005). If a state-level prekindergarten program manager wanted to use assessment data to guide professional development, he or she might insist on analyses that linked the prekindergarten year gains in children's performance to data on the teaching environments the children experienced. So two issues merge in this strategy: having data from well-designed analyses that permit strong conclusions about the program's activities *and* analyses that enable decision makers to make the adjustments they believe will lead to program improvements.

Consider implementing sampling procedures. Assessment systems can be implemented in many ways. As noted in connection with the uses shown in Table 1, a number of uses can be accomplished with data from a sample of programs and children. Thus, one of the major options is either to conduct a census of all children and programs in the state-funded system or to use one of several sampling strategies. If all children are assessed in all classrooms of all programs, it is theoretically possible to provide feedback to programs at the individual child level so that teachers could make instructional decisions. (Whether one would want to do this is a separate issue important to debate—and this debate would largely hinge on the nature of the assessments.) If the accountability system conducts sampling (perhaps with the goal of reducing costs or the

burden on programs, staff, and children), then feedback (and resultant program improvement activities) would be limited to the classroom, center, or program levels, depending on the sampling design. This might be the choice of the state prekindergarten program manager who wants to provide feedback to local programs about the performance of their individual classrooms or children. Data from a sample of programs would meet the needs of the state budget office that is interested in drawing conclusions justifying the budget expenditures.

Sampling requires careful attention to the characteristics of programs and children and how they are distributed across the state. Typically, a sampling statistician would be required to work with the policy staff to determine the most efficient way to draw the sample. Then, to carryout the design, the state would need to work with local programs to obtain class lists and ensure that the information necessary to achieve the desired sample (such as children's race/ethnicity, language, and disability status) is available early in the program year. With sampling, the user would also be well advised to ask that results be presented using confidence intervals so that it is clear what risk there is in accepting the average results as truly reflecting what is going on in the programs.

Sampling also can occur within the assessments by systematically varying the items that individual children are administered. Strategies such as matrix sampling enable a state to obtain data on a much wider range of developmental domains at the *program* level without overly increasing the assessment burden on individual children. Such strategies, however, also make data less usable at the child level. Some would consider this a benefit; others would find it a loss.

The consequences of sampling will interact with the nature of the assessment. If, as in the example of Maryland's school readiness model (Grafwallner 2005), the state implements broad assessments spanning seven domains of learning and employs a portfolio-based assessment

method, then it might be both feasible and desirable to assess all children; the results could be very informative to teachers, who would be able to learn about the progress of individual children. If, however, the accountability system resembles the Head Start NRS (Administration for Children and Families 2005), which has elected to assess more selective domains of development and use a very brief and more traditional testing approach, then it would be preferable not to report individual scores (which is the decision the Office of Head Start has made).

Safeguards

Looking back at the challenges facing those who would implement a state early childhood accountability system, the sometimes complex strategies needed, the negative consequences that could result, and the various assumptions (with all their implications), one might be tempted to avoid this enterprise altogether. But if the potential benefits are real, and if they *can* be accomplished, it is clearly worth the struggles. Fortunately, a number of actions can effectively safeguard the integrity of the accountability system, and we even have examples in efforts already underway in some states.

Plan. There is no substitute for careful planning and anticipating how data will be used and might be misused. Denno et al. (2005) provide a wonderful example from Ohio, where a set of activities takes place, including self-study and external verification, external observation, data verification, data submission, report to community, professional development, and corrective action.

Invest up front. The early decisions that create the system can be crucial. These include selecting appropriate and valid measures, training assessors and observers, and ensuring careful, accurate analyses. Implementing all of these requires a substantial infrastructure of staff and the resources they need to function. Even though Schorr et al. (1994) suggest that results-based

accountability may reduce bureaucracy, no one begins the process with the expectation of saving money, at least not initially. Nevertheless, carefully planned early investments will be less expensive than later revamping of a hastily conceived effort.

Maintain integrity of the assessment process. This is one of the most crucial aspects of any accountability system. Safeguarding against misuse begins with the selection of appropriate measures and extends to ensuring accurate test administration and interpretation of results, with followup that includes strategies to minimize prospects for teaching to the test. Don't rely on any single (or small set of) measures for producing the assessment data for making program decisions. At the same time, have clear guidelines about what is acceptable in terms of the technical properties of the measures used in the system (NAEYC 2003, p. 14).

Use sampling. As noted earlier, sampling children (or matrix-sampling assessment items) provides some safeguards against misusing data obtained on individual children.

Remember the role of context. Establish guidelines for interpreting child outcome data *in the context* of program quality and demographic information. (NAEYC 2003, p. 14). Think through the attribution question.

Plan ahead for professional development. Program staff are the greatest resource states have for ensuring effective prekindergarten programs. Plan for ongoing professional development, poised to respond to data on the pattern of program strengths and weaknesses. As Grafwallner (2005) describes in the state of Maryland's experience, professional development for teachers has been instrumental in guarding against teacher bias in observing and assessing children.

Establish broad support for the system. As the Maryland example shows, the early care and education community, advocacy groups, and the public were involved in reviewing and

discussing the results. Grafwallner attributes the success of the state system in part to this process.

Monitor the accountability system. Effective program managers will keep track of how all the elements of the accountability system are being implemented. One important purpose of this is to minimize the risk of programs finding ways (inadvertently or advertently) to make themselves look better. This could come about if programs somehow, for example, alter their recruitment and enrollment strategies to "cream" the best students or begin teaching to the test.

Build in strategies for system improvement. Not only can states use assessment data for program improvement, the monitoring and tracking of the *accountability system* activities should produce data on its operation so that continuing development can occur over time.

What Might a State Accountability System Look Like? A Hypothetical Illustration

It may be difficult for states to envision how all these issues would play out since no complete models exist. However, elements of such a model can be found to build on. In what follows, I consider program improvement, one of many potential uses illustrated in Table 1, but a very likely use of accountability assessment data. In this illustration, I assume that the user is a state early childhood program manager or director of prekindergarten programs. One of the major implications of this illustration is the extent to which uses of assessment data drive many crucial decisions that will affect the lives of key staff throughout the state prekindergarten system. Following this illustration, I lay out in a more abbreviated form what the scenario might be like for several other users and uses. For this purpose, I have identified eight key steps in the accountability process.

1. **Determine the uses of assessment data the state is most interested in.** This state has decided that, at least for its initial year of implementation, it will focus on *improving classroom practices* as the outcome it wants most to achieve from its new accountability system. State

early childhood department staff identified two specific actions that they could take if the accountability results showed a need for improvement in particular programs: (1) inservice training for teachers in the curriculum areas that assessment data (both child outcome and classroom process) suggest to be weakest, and (2) frequent monitoring of the poorly performing programs to check on whether improvements are occurring during the year. This use requires data both on child outcomes and program quality.

2. Set expectations for children's learning. This hypothetical state has developed a comprehensive set of "preschool learning standards." These standards outline indicators for children's learning and development that the state expects children to be able to demonstrate when they enter kindergarten. This state identified eight broad domains and articulated from one to four specific dimensions within each, which for this purpose are not important except for the fact that the expected outcomes comprise very broad areas of children's learning and development. This state's standards include the following domains:

- Language development
- Literacy
- Mathematics
- Science
- Creative Arts
- Social and emotional development
- Approaches to learning

Physical health and development

3. Decide on whether to assess status or change, or both. Because this state has such a keen interest in program improvement, the decision makers decided that they need to obtain data both on children's progress during their preschool year and on the characteristics of the programs

they attend. The state decides that to have data on children's progress during the year, both fall and spring assessments will be necessary. After grappling with how frequently classrooms need to be assessed, the state decides that one mid-year assessment will be sufficient. This saves money over having to conduct classroom observations twice a year.

4. Set criteria for distinguishing poorly and satisfactorily performing programs. After thinking about various options, this state decides to look at two criteria: (1) gains in children's scores from fall to spring that represent sufficient progress to declare the program to be effective and (2) an absolute level of classroom teacher-child interaction quality that indicates learning environments in which children can be expected to have the opportunity to develop in most of the areas identified in the standards.[7]

5. Build the state infrastructure. Even before the first assessments take place, the state prekindergarten director works with her partners in county agencies and school districts throughout the state to create awareness of the new program, help them understand why accountability is important and how it can benefit the children they enroll, and obtain input for ideas on professional development and improvement of classroom practice. The director establishes a unit within her department to conduct and/or manage the actual data collection. The director then compiles a list of agencies and contractors with a track record in successful technical assistance and begins to identify ones that could be ready to go into action when the assessment results are in. Next, the director identifies key curriculum developers who would have the capacity to work with the T/TA staff to help strengthen any curriculum areas found to need strengthening.

[7] In the actual use, the criteria would be more concrete than in this illustration. This could be done if the actual child outcome and classroom process measures were specified.

6. Decide on child and program measures and assessment procedures. The state director establishes an Assessment Advisory Committee to develop recommendations for assessment measures that will cover all eight of the domains that the state learning standards call for, as well as procedures for assessing classroom learning environments. She recognizes that final decisions will also need to take into account the burden on children and teachers and overall costs, so in her charge to the AAC she asks that they set priorities, assuming that there is time for 60 minutes of direct assessment of the children and 15 minutes per child for teachers to complete behavioral ratings. She asked that both direct child assessments and teacher ratings be considered because she had read that there were no valid and reliable direct assessments of social-emotional development that could be done within this timeframe. Thus, she thought it important to combine direct assessments with ratings by an adult who knows the child well.

For the program process measures, she asks that the AAC take into account her view that technical assistance can be more effective if the assessments provide rather detailed information about instructional practices in the areas of supporting development of language, mathematical concepts, literacy, and social-emotional development, particularly emotion regulation. She also wants to know about specific teacher-child interactions, not just indicators of global quality.

7. Collect the data. After considering its budget, taking into account the concerns that preschool teachers and directors have raised about how much extra work the accountability system will entail, and consulting with a number of experts in the field, the director decides to implement a sampling procedure. Her experts outline a method in which data will be collected from every program in the state, but that sampling will take place within programs based on their size. She is told that data on eight children per classroom is enough to give a reliable index of how well the children are performing on the eight domains. For small programs, all classrooms will be sampled, but for programs that have more than 10 classrooms, 50 percent will be

sampled. She worries that program improvement efforts might be insufficient if data are available on only half the classrooms of the larger programs, but considers two other factors: (1) some 60 percent of the state's programs have fewer than 10 classrooms, so for all these small programs, data will be available for all classrooms; and (2) for the larger programs, the state can systematically sample different classrooms each year, so that over a three-year period, data from all classrooms will be available.

At the same time, the director is convinced by another set of experts that every child doesn't need to be administered every single item in the very long assessment battery that she's told is needed to capture all eight domains. Thus, she decides to implement matrix sampling. She also decides to invest in the necessary hardware and software to use computer-assisted personal assessments to make it easier to implement the matrix sampling while also reducing costs for data entry. Planning for child and classroom sampling and instrument matrix sampling is a major up-front investment, but the director convinces her budget people that this will pay off ultimately.

8. Plan the analysis and reporting. In her final planning step, the director asks the state research and evaluation office to come up with an analysis plan that will allow her to obtain reports that (1) carefully analyze children's progress over their preschool year (fall-spring gains); (2) give her comparative progress indicators by program and classroom within programs, with break-outs for programs serving substantial numbers of children with disabilities, English learners, and children representing the important racial-ethnic groups in the state; (3) show any progress in the context of the key process dimensions from the classroom observation data; and (4) produce all these analyses while controlling for children's entering status. The director asks for an analytic approach that will report both progress in the preschool year and status at the end of the year so her department can take both into account in planning T/TA for program

improvement. The director expects that the results of the fall-spring analysis will be available by mid-summer following the prekindergarten year. This will enable her to set the corrective actions into motion. With the data decisions she has made, she is able to target the training on those specific teachers whose classrooms (1) show inadequate teaching practices in the classroom observation data *and* (2) produce inadequate progress in children's learning. The director also plans to institute more-frequent monitoring in the identified programs during the following year, and then be able to check their results at the end of the second year.

We could create a number of other illustrations of how these eight steps would play out for different uses and users at the state level, as begun in Table 2 (next page). In this table, state 1 is the scenario just described. States 2 through 4 illustrate the different decisions that might be made if the states had different purposes.

Conclusions

The discussion in this paper suggests that states should employ strategies that will ensure meeting the challenges of implementing statewide prekindergarten accountability systems and consider the safeguards that, together, will lead to an effective system that meets the needs of most stakeholders and has the greatest potential for supporting programs that benefit all children. By way of conclusion from this discussion of a wide range of issues and considerations, I emphasize five sets of actions for state accountability systems:

1. Agree on intended uses with stakeholders and decision makers at the outset, before selecting assessments measures and assessment procedures—this means involving the key potential users in system design.
2. Ensure that the infrastructure is in place to support the necessary training, assessment, analysis, and reporting.
3. Anticipate system improvement needs and have procedures in place for implementing them when results show the need.

4. Anticipate the challenges to creating an effective system so that misuses of data are minimized and appropriate uses are maximized.

5. Build in all the safeguards needed to avoid unintended consequences and maximize the probability of an effective accountability system.

Table 2. Illustration of how decisions about uses drive many other decisions in the state prekindergarten accountability system

Step in Developing the State's Accountability System	State 1: User Is State Early Childhood Program Manager or Director of Prekindergarten Programs	State 2: User Is State Early Childhood Program Manager	State 3: User Is State Legislative Committee with Oversight of Early Care and Education Programs	State 4: User Is State Office with Oversight of Social Programs, Including Prekindergarten
1. Determine most important use of the accountability system	Improve classroom practice in poorly performing programs	Convince legislature that prekindergarten programs are "working"	Decide which programs should continue receiving funds, which should be put on probation, which should be defunded.	Improve classroom practice across the state in areas deemed most in need of improvement
2. Set expectations for children's learning	Preschool learning standards that span eight domains of development	Preschool learning standards spanning three domains of development	Preschool learning standards that span four top-priority domains of development: literacy, math, emotion regulation, and health status (because of concerns about childhood obesity)	Legislative priorities for "readiness" in language, literacy, and mathematics
3. Decide whether to assess status, change, or both	Both status at the end of preschool and fall-spring growth in prekindergarten	End of prekindergarten status only	Entering kindergarten status only	Growth from fall to spring during the prekindergarten year
4. Set criteria for distinguishing poorly and satisfactorily performing programs	Fall-spring gains showing progress indicative of an effective program, along with high teacher interaction quality	Overall average status above predetermined cut score in the three domains	Average status of children at the beginning of kindergarten with three levels: satisfactory, needs improvement, and inadequate	Average fall-spring growth at each program in the three domains with 2 levels: adequate growth and needs improvement

Step in Developing the State's Accountability System	State 1: User Is State Early Childhood Program Manager or Director of Prekindergarten Programs	State 2: User Is State Early Childhood Program Manager	State 3: User Is State Legislative Committee with Oversight of Early Care and Education Programs	State 4: User Is State Office with Oversight of Social Programs, Including Prekindergarten
5. Build infrastructure	Build relationships with programs and school districts, create awareness, plan data collection, prepare T/TA contractors, plan analyses	Build relationships with programs and school districts, create awareness, plan data collection, plan analyses	Build relationships with programs and school districts, create awareness, plan data collection, plan analyses	Build relationships with programs and school districts, create awareness, plan data collection, prepare T/TA contractors, plan analyses
6. Decide on child and program measures and assessment procedures	Both direct child assessments and teacher ratings that measure all 8 domains; classroom observation systems that tap into key teaching behaviors	Direct child assessments that cover the 3 domains	Child outcome assessments only in 4 areas: literacy and math (direct assessment), emotion regulation (rated), and health status (height and weight because of concerns about childhood obesity) .	Direct child assessments that cover the 3 domains
7. Collect the data	Assess sample of children in all classrooms in all programs	Assess sample of children in a sample of classrooms in a representative sample of programs	Assess sample of children in all programs; no classroom process assessments	Random sample of children in representative sample of programs throughout the state, with sample size that's necessary only for reliable program-level averages
8. Plan analysis and reporting	Complex analysis that examines gains in each local program, controlling for classroom processes and children's entering status; provide both status and change scores.	Produce one score for each domain that accurately reflects status of preschool children in spring of their preschool year in each local program	Entering kindergarten status only for each local program. Cut scores that will lead to decisions about funding.	Program-level; scores that show degree of child growth during preschool in each local program

References

Administration for Children and Families (2005, February). *Assessment of children in Head Start programs: Report to Congress.* Washington, DC: US Department of Health and Human Services.

Administration on Children, Youth and Families. (2000). *Head Start Child Outcomes Framework.* Washington, DC: Head Start Bureau.

Christina, R., & Nicholson-Goodman, J. (2005). *Going to scale with high-quality early education: Choices and consequences in universal pre-kindergarten efforts.* Santa Monica, CA: RAND Corporation.

Datta, L.E. (1976). The impact of the Westinghouse/Ohio evaluation on the development of Project Head Start: An examination of the immediate and longer-term effects and how they came about. In C.C. Abt (ed.). *The evaluation of social programs.* Beverly Hills: Sage, pp. 129-191.

Denno, D., Miller, S., & Joyce, E. (2005). Supporting school success: Ohio's early learning accountability system. Paper prepared for the National Task Force on Early childhood Accountability.

Gilliam, W.S., & Marchesseault, C.M. (2005). *From capitols to classrooms, policies to practice: State-funded prekindergarten at the classroom level: Part 1. Who's teaching our youngest students? Teacher education and training, experience, compensation and benefits, and assistant teachers.* Report prepared for the National Institute for Early Education Research and the Pew Charitable Trusts. New Haven: Yale University.

Goldschmidt, P., Roschewski, P, Choi, K., Auty, W., Hebbler, S., Blank, R., & Williams, A. (2005, October), *Policymakers' guide to growth models for school accountability: How do accountability models differ?* Washington, DC: Council of Chief State School Officers.

Government Accountability Office (2005, May). *Head Start: Further development could allow results of new test to be used for decision making.* Report GAO-05-343. Washington, DC: Author.

Grafwallner, R. Maryland model for school readiness (MMSR) kindergarten assessment: A large-scale early childhood assessment project to establish a statewide instructional accountability system. Paper prepared for the National Task Force on Early childhood Accountability.

Iruka, I.U., & Carver, P.R. (2006). *Initial Results from the 2005 NHES Early Childhood Program Participation Survey* (NCES 2006-075). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Karoly, L.A., & Bigelow, J.H. (2005). The economics of investing in universal preschool education in California. *RAND Corporation Monograph Series.* Santa Monica, CA: RAND Corporation.

Linn, R.L., Baker, E.L, & Betebenner, D.W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3-16.

Maxwell, K.L., & Clifford, C.M. (2004). School readiness assessment. *Beyond the Journal: Young Children on the Web*, January. (http://www.naeyc.org)

Meisels, S.J., Atkins-Burnett, S., Xue, Y., Bickel, D.D., & Son, S.H. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives, 11*(9). http://epaa.asu.edu/epaa/v11n9/ downloaded 8/1/05.

National Association for the Education of Young Children. (2002). Early learning standards: Creating the conditions of success. (Joint position statement of the NAEYC and the National Association of Early Childhood Specialists in State Departments of Education.) Washington, DC: NAEYC.

National Association for the Education of Young Children. (2003). Early childhood curriculum, assessment, and program evaluation: Building an effective, accountable system in programs for children birth through age 8. (Position statement adopted November 2003.) Washington, DC: NAEYC.

National Governors' Association. (2003). Aligning the goals of preschool and K-12. In *Governors' Forum on Quality Preschool, Issue Brief Series.*

Ravitch, D. (2005). Every state left behind. *New York Times*, November 7.

Rivlin, A.M. (1974). *Social policy: Alternate strategies for the federal government.* Washington, DC: The Brookings Institution. General Series Reprint 288.

Rosenthal, E., Rathbun, A., & West, J. (2005). Regional differences in kindergartners' early education experiences. *Education Statistics Quarterly, 7*(1,2), 15-24.

Schorr, L., Farrow, F., Hornbeck, D., & Watson, S. (1994). The case for shifting to results-based accountability. Washington, DC: Harvard Project on Effective Services.

Scott-Little, C., Kagan, S.L, & Clifford, R.M. (2003). *Assessing the state of state assessments: Perspectives on assessing young children.* Greensboro, NC: SERVE.

Scott-Little, C., Kagan, S.L., & Frelow, V.S. (2003, June). *Standards for preschool children's learning and development: Who has standards, how were they developed, and how are they used?* Greensboro, NC: SERVE.

Shepard, L., Kagan, S.L, & Wurtz, E. (1998). *Principles and recommendations for early childhood assessments.* Washington, DC: National Education Goals Panel.

Stipek, D. (2005). Early childhood education at a crossroads: Access to preschool has come a long way, but critical choices lie ahead. *Harvard Education Letter*, July/August 2005. http://www.edletter.org/past/issues/2005-ja/crossroads.shtml Downloaded 1/10/06.