Herr Research Center for Children and Social Policy
Erikson Institute

Number 6

Occasional **Paper**

Accountability in Early Childhood: **No Easy Answers**

Samuel J. Meisels

Executive summary: Politicians, policymakers, journalists, and scholars want to know that taxpayer-supported programs for young children work. Indeed, accountability has become the centerpiece of federal education policy, and states have been quick to follow suit. Yet increasingly, the measure of accountability—whether or not a particular program works—has been reduced to how well a young child performs on a mandated test. High-stakes → continued

About the author

SAMUEL J. MEISELS is one of the nation's preeminent researchers on developmental assessment in early childhood. President of Erikson Institute since 2002, Meisels is also professor and research scientist emeritus at the University of Michigan, where he taught for 21 years. He holds a bachelor's in philosophy from the University of Rochester and an M.Ed. and Ed.D. in education from Harvard. Widely published, Meisels is coeditor of the *Handbook of Early Childhood Intervention* (Cambridge University Press, 2000), and coauthor of *Developmental Screening in Early Childhood: A Guide* (NAEYC, 2005) and *The Work Sampling System* (Pearson Early Learning, 2001). He is former president of the board of directors of Zero to Three: The National Center for Infants, Toddlers, and Families and is an adviser to the national Head Start Bureau. A version of this paper is forthcoming in R.C. Pianta, M. J. Cox, & K. Snow (Eds.), *School readiness, early learning, and the transition to kindergarten.* Baltimore: Paul H. Brookes.

Occasional Paper | Number 6

March 2006

Herr Research Center
for Children and Social Policy
Erikson Institute

Occasional Papers are issued periodically in conjunction with the semiannual newsletter *Applied Research in Child Development,* published through the Herr Research Center for Children and Social Policy at Erikson Institute. The preparation and dissemination of Herr Research Center publications is made possible through generous grants from the Joyce, McCormick Tribune, and Spencer Foundations.

03-06/5M/PN/05-289/Design: Sorensen London, Inc.

→ from the cover decisions, including continued program funding, employment and pay of teachers, and student retention, are being made on the basis of this single data point.

How did a testing approach originally developed for middle- and high-schoolers come to be applied to very young children? Are the results of such tests reliable, and if they are, can a narrow range of information about a single child at a particular time be used to evaluate teaching or curriculum? In this paper, written as a chapter in the forthcoming *School Readiness, Early Learning, and the Transition to Kindergarten* (R.C. Pianta et al., Eds.), Samuel J. Meisels examines the genesis of accountability testing in preschool and refutes the quality-assurance, production-model assumptions that underlie its use with young children. Citing the best available research, he summarizes the arguments against such testing in early childhood: the practical problems of measuring the developmentally unreliable; unintended but real consequences for teaching and learning; the failure of such tests to account for tremendous differences across the preschool population in prior opportunities to learn; and the demonstrably weak association between academic/cognitive measures in preschool and like measures in first and second grade. Meisels goes on to examine how each of these facts or circumstances contributed to the failure of Head Start's National Reporting System, one of the largest-scale examples of early childhood accountability testing to date.

Finally, Meisels takes up the question of how to measure program effectiveness and program quality. He argues for program evaluation: collecting data on structural and dynamic characteristics of programs (child-staff ratios, staff training, developmentally appropriate practice, positive interaction between children and staff, parental involvement, etc.), key demographic variables, and finally, programs' impact on children. To measure the latter, Meisels proposes creating an assessment based on item response theory (IRT), using a metric that describes children's' relative position on a developmental path. Such an assessment will not only indicate whether children are learning. It will enable the analysis of program elements, pedagogical techniques, and child outcomes to determine whether particular aspects of a program or child and family background are more or less strongly associated with child outcomes.

By learning what works for whom, we can move beyond simply identifying a particular program's outcomes to determining what we can do to help that program—and the young children it serves—succeed.

Accountability in early childhood: No easy answers

Education and social service programs in the first part of the 21st century are dominated by accountability. Publicly at least, politicians, policymakers, journalists, and scholars are focused on outcomes—on what works. For the U.S. Department of Education, this vow to hew to the path of accountability has even been translated into law. President George W. Bush's signature education legislation, the 2002 reauthorization of Title I, entitled the *No Child Left Behind Act* (NCLB), made accountability the centerpiece of educational policy and test scores the sole means of demonstrating it. Annual testing in reading and math is required for grades 3–8 and severe consequences, leading even to closures for schools not making "adequate yearly progress" as shown by scores on standardized tests, are spelled out in the legislation. Despite the problems inherent in this law (see Lynn, 2005), including "perverse incentives" that result in *lowering* rather than raising standards (Ryan, 2004), NCLB has dominated educational practice since its passage. Nearly all discussion of school reform, curriculum models, and novel school governance structures (other than charter schools) has given way to a single-minded attempt to increase students' scores on high-stakes tests.

Programs for children enrolled in preschool and the early elementary grades are also affected by accountability pressures. Some states have instituted annual kindergarten accountability testing. Others are attempting to link testing in kindergarten to performance of state-funded pre-K programs during the previous year. The most extensive use of high-stakes testing has taken place in Head Start, where a twice-yearly standardized test—the *National Reporting System* (NRS)—was first administered in 2004.

In the face of this near-obsession with accountability, educators and policymakers have sought expedient solutions to the complex problems of determining who has learned what, how much they learned, and how well they learned it. Conventional norm-referenced tests enable us to rank and order individuals according to a single, easily understandable metric. But their closed-ended questions do not measure children's natural curiosity, ability to solve problems, or emergent creativity. They are unable to describe individual patterns of learning and teaching; they do not give voice to cultural and ethnic differences that may depart from the mainstream; and they have become vested by

our educational system with disproportionate power over teachers' decisions regarding curriculum and the utilization of instructional time.

As test scores begin to be used for high-stakes purposes, they are increasingly viewed not as one datum about student performance, or one source of information about student learning, among many. Rather, they are perceived as sufficient evidence to render decisions about retention, promotion, teachers' expertise, and school success. These are the consequences that are typically associated with high-stakes testing (Madaus, 1988), despite the fact that it is well-known that important educational decisions should be based on multiple sources of information (Heubert & Hauser, 1999). Because of the limited range of information commonly sampled by high-stakes tests and their closed-ended questions and responses, they can distort the educational process by suggesting that one indicator of learning can stand for the whole of learning (Corbett & Wilson, 1991). In this type of a results-oriented framework, teaching becomes preparation for testing.

Some commentators have gone so far as to say that instruction that is primarily test-oriented is "anti-educational" (Parini, 2005, p. 10). Such teaching is viewed as "a kind of unpleasant game that subverts the real aim of education: to waken a student to his or her potential and to pursue a subject of considerable importance without restrictions imposed by anything except the inherent demands of the material." The test-driven perspective may take its greatest toll on young children who have not yet learned to play the "school game." For them, an early introduction to high-stakes testing may influence their long-term attitudes not just about what takes place in schools, but about their overall academic capabilities and their sense of self. With the expansion of large-scale testing to preschool and the first few years of formal schooling, it is essential that we explore the implications of applying a testing paradigm designed for older students to those younger than age eight.

My purpose here is four-fold. First, I will focus on the reasons behind the growth of accountability testing in preschool and the early grades. Then, I will explain why accountability testing is such a problematic activity in the first eight years of life. Next, I will illustrate the content and rationale for these tests by using the NRS as an example. Finally, I will explore other means of responding to the major questions that policymakers expect high-stakes testing to provide by discussing the parameters of program evaluation. In this fourth section I will also introduce the elements of a potentially less problematic design for outcome assessment of young children.

What Policymakers Want to Know About the Effectiveness of Early Childhood Programs

Early childhood care and education as we know it today does not have an extremely long history. As detailed elsewhere (Bowman, Donovan, & Burns, 2001; Lazar & Darlington, 1982; Meisels & Shonkoff, 2000), the first public kindergarten programs did not appear before the mid-19th century, and research-based programs for children younger than age five did not begin until the early 1960s.

The model programs of the 1960s sought to obtain evidence about how preschool could reverse the "cycle of poverty" that led to poor education, poor job prospects, and poor parenting (Farran, 2000; Halpern, 2000; Zigler & Valentine, 1979). Reflecting the knowledge base of that time, research sought to link the effectiveness of these programs to growth in IQ scores of poor children, this seen as a first step in changing the life chances of these children (Schweinhart & Weikart, 1980). The psychologist Urie Bronfenbrenner reviewed the data from these programs in a 1974 monograph entitled *"Is Early Intervention Effective?"* His question continues to be posed today, regardless of how many times it has been answered or how often it has been reformulated (see Meisels, 1985; Shonkoff & Phillips, 2000).

Bronfenbrenner's (1974) view was that the family is the most efficacious and economical system for fostering and sustaining the development of the child. Involvement of the child's family as active participants is critical to the success of any intervention program, and without family involvement the effects of intervention erode quickly. Involvement of parents has the potential for establishing an ongoing system that can reinforce the effects of a program and that can help sustain them after the program ends. Thus, the family appears to be a key target on which to focus intervention efforts.

But Bronfenbrenner's review, while ahead of its time in its focus on the child as part of a system or network and thus foreshadowing Bronfenbrenner's landmark work on the ecology of human development (Bronfenbrenner, 1979), fell victim to the implied view that children, families, and interventions are relatively homogeneous and uniform. Hence, the title of his monograph, *"Is Early Intervention Effective?"* rather than, Are Early Interventions Effective? We have learned since then that we must ask not one question but many. The task is not to find the best intervention for everyone; the goal is to determine the best intervention for this child and family at this time and in this situation.

Policymakers today seem to be ensnared in the same fallacy of searching for uniform solutions to disparate problems. But the questions we hear from them today are somewhat different from Bronfenbrenner's. One issue that emerged at the beginning of the 1990s, when the National Goals Panel was active, is derived from the first national goal that all children will be ready for school by the year 2000 (Kagan, 1990). Stated simply this question is, Are children ready for kindergarten? However, the Goals Panel resource groups that dealt with this question in the '90s made it clear that there is no simple answer to this question. Instead of suggesting a common set of skills that all children must master in order to be considered "ready"—skills which the familiar technology of testing could readily evaluate—scholars noted that children's differing early experiences and heterogeneous cultural and familial environments render a single test at the outset of school misleading at best (Kagan, Moore, & Bredekamp, 1995; Meisels, 1999). Indeed, as the Goals Panel groups noted, early childhood development is multifaceted. It includes the domains of cognition and general knowledge, language and literacy, motor and physical development, socio-emotional development, and approaches to learning. In short, the "readiness question" cannot be answered easily or quickly, despite policymakers' pressing need for information about how well children are doing in school.

In recent years, public support for pre-kindergarten programs has grown dramatically. Previously, the largest public investment in pre-K programs was the federal Head Start program, which today serves more than 900,000 children at a cost in excess of $7 billion. But over the past 10 years, state pre-K programs have grown to where they nearly match Head Start in terms of number of children served (almost 750,000 [National Center for Early Development and Learning, 2005]), though the amount of money spent ($2 billion in 2002 (Stipek, 2005]) is lower than Head Start, in part because pre-K programs rely on so many in-kind contributions from local school districts and other sources that are difficult to tabulate. Pre-K programs are now offered by 43 states and the District of Columbia, and more than 10 states either have or are exploring the option of providing universal pre-K services (NIEER, 2005). With the growth of these programs, the "readiness for kindergarten" question has been sharpened and expanded. Now policymakers are asking two questions that are corollaries of their earlier readiness query: Are children learning? and, Are public funds being used wisely?

These two questions are extremely important, though the methods used to obtain meaningful answers to them are not obvious. The first question about children's learning goes to the heart of why policymakers have embraced pre-K and other early childhood programs. The U.S. is facing a prolonged and pernicious achievement gap between white and non-white students and between students from economically more advantaged vs. less advantaged families. With the majority of parents now in the workforce, safe and sound child care is not merely a necessity to maintain our economy. The pre-K "solution" is meant to have an impact on these fundamental inequities. Like the early model preschool programs of the 1960s and the original formulation of Head Start from that same era, pre-K programs today are intended to close the gap, equalize opportunity, and enable our society to derive benefits from and for more of its citizens. It is no wonder that policymakers are becoming impatient for answers.

This analysis leads directly to the second question, regarding the value of the public investment that these programs represent. The reasoning goes something like this: If the programs are not improving learning—if they are not closing the achievement gap—then how can we justify their cost?

Some might argue that numerous programs are supported by public dollars without proof of their efficacy (e.g., public parks, civic holiday decorations, or even some contemporary birthing innovations, to select just a few examples). Others could point out that the expectation that program efficacy continue to be demonstrated even when evidence of effectiveness has been shown before could be a higher standard than is required for other professions. For example, physicians and other health professionals have a great deal of leeway in how they implement interventions for their patients, as long as they follow an established protocol. Similarly, early care and education has a growing list of efficacious experiments and implementations that provide the basis for the work of pre-K professionals (Brooks-Gunn, Fuligni, & Berlin, 2003; Karoly et al., 1998; Meisels & Shonkoff, 2000; Shonkoff & Phillips, 2000). But there is a difference. Health care professionals are all trained to a particular level of recognized and acceptable expertise and their working conditions generally enhance their professional growth and expertise. Most early childhood professionals do not have this background or supportive professional environment (see Hart & Schumacher, 2005). Furthermore, the fact that children differ so greatly from one another in their early experiences, opportunities to learn, genetic

inheritance, and family structure, among other variables, only adds to the challenge of evaluating early education outcomes.

In short, the two questions policymakers are asking are reasonable and appropriate. Our debate is not about the questions, though if other changes were made in the preparation and working conditions of early care and education professionals, it is possible that the insistence on obtaining answers for each local or state situation might diminish. The problem is that the high-stakes methods being proposed to determine if a program is or is not effective are themselves open to question regarding their accuracy, appropriateness, and meaningfulness.

The Arguments Against High-Stakes Testing in Early Childhood

High-stakes testing refers fundamentally to the uses made of test scores, rather than to any particular test or type of test data (Madaus, 1988; Mueller, 2002). To the extent that test information, or any other type of comparative data, is used to make decisions about who should receive rewards or experience sanctions, then that test is considered high-stakes. In early childhood, rewards can take the form of public attention, additional funds for teachers or materials, increased salaries, or improved facilities. Sanctions include holding children back or enrolling them in extra year programs, wresting control of curriculum from teachers, or even program closure (Meisels, 1992).

Although high-stakes testing is common in the K–12 world, it is less frequently encountered in early childhood. Previous examples include the widespread use of the Gesell School Readiness Test to determine whether children could enter kindergarten (Shepard & Smith, 1986), and the statewide adoption of an adapted form of the California Achievement Test to decide if kindergarten children could be promoted to first grade (Meisels, 1989). Of course, the incentive structure of NCLB for third–eighth graders is built entirely around high-stakes testing, with the ultimate sanction being closure of a poor performing school (Ryan, 2004).

Many scholars have expressed misgivings about the use of high-stakes tests as a means of determining a program's overall achievement level (Madaus & Clarke, 2001). Some even claim that it is "scientifically indefensible" to use the average achievement scores of a school to judge how well a school is performing (Raudenbush, 2005). Raudenbush points out that "If you want to measure what goes on in a school, you have to develop measures that look at

the educational process and practices, not just at children's relative achievement" (ibid.). Conventional high-stakes tests do not measure the quality of the educational practices at a particular school or children's relative rates of learning.

The problems of using high-stakes tests with young children are even more severe. Four reasons stand out for not using high-stakes tests with young children (see Meisels, 1994, and Meisels & Atkins-Burnett, 2006, for a discussion of these and other related points).

Practical Problems of Measurement. Young children are developmentally unreliable test takers. They have a restricted ability to comprehend such assessment cues as verbal instructions, aural stimuli, situational cues, or written instructions. Further, questions that require complex information-processing skills—giving differential weights to alternative choices, distinguishing recency from primacy, or responding correctly to multistep directions—may cause a child to give the wrong answer. In addition, young children may not be able to control their behavior to meet the demand characteristic of the assessment situation—whether this is because they are affected by fatigue, boredom, hunger, illness, or anxiety, or simply because they are unable to sit still and attend for the length of time required.

Unintended Consequences. High-stakes tests may result in long-term negative consequences for young children. This follows because we know that the structure of teaching and learning can be affected negatively by focusing on test results, thus resulting in measurement-driven instruction, which can homogenize what might otherwise be a very heterogeneous curriculum. Also included are the potential negative effects on children's sense of self-worth and self-perception that judgments based on test results can convey to them. Rist (1970) described these effects in great detail by noting how both teachers and children were changed by what Rosenthal and Jacobson (1968) called the "Pygmalion effect," when teachers' perceptions are altered by information from tests and other sources external to the classroom, regardless of their accuracy. For young children, the risk is that children will feel stigmatized and be tracked into low achieving groups that will further confirm their sense of powerlessness and limited potential. Their estimates of their own abilities—their self-perceptions and their motivation and ultimately their achievement—are likely to suffer as a result.

Opportunity to Learn. Children's opportunities to learn differ greatly in early childhood, and no period of common schooling (such as occurs to some extent in K–12) is available to them. "Opportunity to learn" concerns what children have been taught before entering the program in which they are enrolled. The range of opportunities to learn in early childhood mirrors the fundamental differences in society and especially reflects the challenges faced by poor and disadvantaged children prior to even arriving at the school door. To assume that a test administered at the outset of school can be used to make valid predictions that may have long-term consequences, is to believe that these inequities are virtually immutable. The task of schooling is to begin to overcome these inequities by providing an environment in which children can learn what they have not yet been taught and can begin to achieve. If we ignore differences attributable to opportunity to learn, as conventional accountability measures do, we are begging the fundamental question of what individual children need and how we can fashion a curriculum that is responsive to these needs.

Variability and Predictability. The final argument for not using high-stakes testing in early childhood derives from the extensive variability and change that marks early development. LaParo and Pianta (2000) documented this instability of development in a meta-analysis of 70 longitudinal studies. Their purpose was to study the associations between academic/cognitive and social/behavioral measures in preschool and kindergarten with like measures in first and second grade. They found that only about a quarter of the variance in early academic/cognitive performance was predicted by preschool or kindergarten cognitive status; only 10% or less of the variance in K–Grade 2 social/behavioral measures was accounted for by similar measures at preschool or kindergarten. LaParo and Pianta conclude that "instability or change may be the rule rather than the exception during this period" (p. 476). In short, their study shows that tests used to make predictions—even relatively short-term predictions—are insufficiently stable to justify assigning stakes based on them. Given that young children are undergoing significant changes in their first eight years of life in terms of brain growth, physiology, and emotional regulation, and recognizing that children come into this world with varied inheritance, experience, and opportunities for nurturance, it is not difficult to imagine that a brief snapshot of a child's skills and abilities taken on a single occasion will be unable to capture the shifts and changes in that development. To draw long-term conclusions from such assessments seems baseless.

Additional research that supports this view is put forward by Kim and Suen (2003). Using hierarchical linear modeling they report a validity generalization study of 716 predictive correlation coefficients from 44 studies. Their purpose was to determine if the predictive validity coefficients of early assessments could be used to draw generalized conclusions about later achievement or success in school. The authors posed two questions. First, is it possible that "predictive validity is unique to each early assessment procedure and unique to each specific set of local testing conditions" (Kim & Suen, p. 548)? This would be the case if predictability was affected by sample characteristics, length of time between prediction and outcome, or the outcome criterion itself. Their second question focused more specifically on statistical artifacts: Are there statistical or measurement errors that potentially prevent us from obtaining reliable predictions of outcomes from early childhood assessments? The errors or statistical artifacts include a range of variabilities concerned with test and criterion unreliabilities, local restricted ranges of scores, and other sampling errors.

Their study answered these questions definitively. They demonstrate that predictive validity coefficients in early childhood are different in different situations. Stated in another way, they point out that "the predictive power of any early assessment from any single study is not generalizable, regardless of design and quality of research. The predictive power of early assessments is different from situation to situation" (p. 561). This does not mean that there are no early childhood tests with predictive value. Rather, Kim and Suen's study demonstrates that predictions from early childhood assessments cannot be generalized meaningfully. Even if we were to average all adjusted predictive coefficients in order to obtain a "typical" overall prediction, this could give misleading information. This follows because an overall average coefficient conceals unaccounted-for variation and is not therefore representative or meaningful.

Kim and Suen have shown that if you use a test to demonstrate predictive validity in one situation, and another test in a different situation, it is unjustified to assume that the same thing is being measured in these situations or by these tests. Each assessment and each set of conditions needs to be treated as unique. However, this does not imply that individual outcome studies are invalid. Rather, this study contends that early assessments *in general* are not predictive of future performance.

In brief, both Kim and Suen's and LoParo and Pianta's studies arrive at conclusions that are very similar, although they get there by different means.

They help us to understand the consequences of developmental instability in early childhood development and they remind us that tests of accountability that overlook this variability have a high likelihood of providing unsubstantiated conclusions. We will now turn to an account of a national test of Head Start children that appears to incorporate nearly all of these problems.

A Failed Experiment: The National Reporting System

A milestone in U.S. educational history took place in the fall of 2003. That year the largest administration of a single standardized test—the Head Start National Reporting System, or NRS—was launched. At an estimated total cost in excess of $25 million annually (including direct and indirect costs), approximately 450,000 4-year-olds from every state and nearly every locale in the nation began to be administered the NRS twice yearly. This may be the largest test administration in U.S. history. Even the NAEP, or *National Assessment of Educational Progress*—known as the "Nation's Report Card" (Pellegrino, Jones, & Mitchell, 1999) included no more than 350,000 students in its 2005 administration. Moreover, the NAEP uses a matrix sampling approach, so that different students receive different parts of the test at different times. Ultimately, the scores from these separate administrations are combined statistically to provide an overview of the nation's school performance.

The NRS is different. All Head Start children aged four and older who speak English or Spanish are administered the entire test twice yearly. The stated purpose of the test is three-fold: (1) to enable programs to engage in self-assessment and improvement; (2) to target needed training and technical assistance efforts; and (3) to monitor programs' performance in order to determine if public funding should be continued (Administration for Children and Families, 2003).

The test is a classically "top down" policy initiative that high level government officials directed HHS and ACF bureaucrats to put it in place post haste. The decision to create such a test was announced less than a year before it was implemented. In only nine months it was developed and piloted on a small number of children and programs, 30,000 teachers or their surrogates were trained, and the test was manufactured and sent out to the field. This probably set a record for a national assessment. For example, the assessments that became part of the 22,625-child Early Childhood Longitudinal Study-Kindergarten cohort required more than three years of development, piloting, and extensive field

testing and analysis before they were considered ready for widescale use (West, Denton, & Germino-Hauskin, 2000).

When the NRS was announced, many in the field urged that, if the test had to be given, only a sample of children in Head Start be tested. But the HHS administrators wanted to test the population, not a sample. The reason given for this was that without testing every child in every program it would be impossible to answer the efficacy, or accountability, question about those programs. In short, HHS wanted answers to the two questions raised earlier: Are children learning? and Are public funds being used wisely?

HHS eventually funded a small-scale evaluation of the test, but very little oversight was devoted to the preparation or implementation of the assessment. A Technical Work Group, to which I was appointed, was charged with advising the contractor who developed the test and the government officials who had responsibility for implementing it. But the Work Group was not given an opportunity to review the test items before they went to the field in the fall of 2003.

To some extent, the test items are a parody of a well-developed standardized test. Despite the fact that during a visit to a Virginia Head Start program in July 2003 President Bush said that "we would be defeating the purpose of accountability before we even began it if we … give standardized tests to 4-year-olds" (White House Press Release, 2003), the methodology used here is not much different from that employed in NCLB. Although individually administered, the NRS is fundamentally a high-stakes test that relies extensively on multiple-choice items. The test is composed of five subtests, including two language screeners to determine if the child is English- or Spanish-speaking, and tests of vocabulary (derived from the *Peabody Picture Vocabulary Test)*, letter-naming skills, and early math skills.

Much has been made of the culture- and class-specific nature of the vocabulary chosen from the Peabody Picture Vocabulary Test—such words as swamp, vase, awarding, and horrified—and of the problems with the Spanish language test (see Meisels & Atkins-Burnett, 2004). Also of great concern is the linguistic burden and psychometric construction of the math items that assume that Head Start 4-year-olds can attribute causality, do subtraction, use standard metric units, and understand the subjunctive case. Even the letter naming task on the test is misconceived and reflects a lack of understanding about what rapid letter naming teaches us about young children's skills in early literacy. These problems were not corrected throughout the life of this test, though because of public

outcry, Congressional complaints, and prodding from members of the Technical Work Group, they became less egregious.

In May 2005 the problems with the NRS were highlighted in a report to Congress by the U.S. General Accountability Office (GAO, 2 005). In a monograph entitled, "*Further Development Could Allow Results of New Test to Be Used for Decision Making*," the GAO concluded that

> As of February 2005, [the] Head Start Bureau had not conducted certain analyses on NRS results to establish the validity and some aspects of the reliability of the assessment . . . . The NRS by itself does not provide sufficient information to draw conclusions about the effects of Head Start grantees on children's outcomes—information that would support use of the NRS for Head Start grantee accountability. (p. 23, 26)

In short, after a year of study which included 12 site visits to Head Start programs in five states, a review of data and documents, interviews with multiple informants, and advice from three national experts, the GAO found the NRS to produce data that are suspect and to have potentially harmful unintended consequences. As the report notes,

> There is a concern that local Head Start programs will alter their teaching practices and curricula based on their participation in the NRS . . . . [A]t least 18% of grantees changed instruction during the first year to emphasize areas covered in the NRS. (pp. 19-20)

High-stakes tests—and although not yet used for high-stakes purposes, the NRS was designed, among other things, to be such a test—change instruction. They narrow the range of opportunities to learn to those included on the test, even if the child could learn more effectively with a different approach or different content. This problem accompanies such tests at all levels of administration (Herman, 2004; Johnson & Johnson, 2002; McNeil, 2002). In early childhood, however, and in particular, in Head Start, high-stakes testing may have a more pernicious effect than among older students.

Because the Head Start workforce contains fewer than 30% bachelors-prepared teachers and, as of 2003, only 27% who hold even an associates degree (Hart & Schumacher, 2005), it is likely that many teachers in Head Start will alter their teaching to conform to the pedagogical model implicit in this test. Without more training, these teachers will not be able to critically analyze what

is being asked of them and their children and make allowances for individual differences; this state of affairs was documented by the GAO report.

As is the case with other high-stakes exams, the NRS implies a model of pedagogy (Elmore, 2004; Kornhaber & Ornfield, 2001). It is a model of passive reception, of pouring into a vessel knowledge and skills that are needed for competence, rather than recognizing learning as active and teaching as a joint process of interaction between child and adult. An active view of learning, fundamentally based on enhancing relationships between teachers, children, and challenging materials, is nowhere to be seen in this test, although Head Start has been committed to a constructivist outlook on teaching for many years (Zigler & Muenchow, 1992). Yet, when you know that the results of a test will be used to make decisions that may affect your program's continuation and other things you value, you are sorely tempted to begin teaching to the test. Not only does this lead to a great deal of what is called measurement-driven instruction (Darling-Hammond & Rustique-Forrester, 2005) between the fall and spring administrations, it also raises the possibility of "gaming" the system by arranging for children to score low in the fall and then make marked progress by the spring. After all, most of the testing is done by teachers or others who are part of the program and who themselves will be affected by the NRS scores. The potential impact of this test on 3-year-olds as their teachers spend a year preparing them to identify vocabulary words, name letters, and solve counting and measuring tasks can also not be overlooked. In devoting their time and energy to preparing children to perform well on the tests for 4-year-olds, teachers may be ignoring many other elements of learning that are critical for acquiring more advanced skills later on.

In brief, the pedagogical model implicit in the test is highly questionable for young children. But an even more invidious problem emerges from the overall rationale for the NRS. This rationale is associated with the discussion of the achievement gap mentioned earlier.

Policymakers in Washington and elsewhere have long recognized that poor children, and in particular, children enrolled in Head Start, do not start school with skills equivalent to those from more affluent backgrounds (Haskins & Rouse, 2005; Lee & Burkham, 2002). As noted earlier, these policymakers believe that if Head Start were doing its job, this discrepancy—this incipient achievement gap—would be eliminated. This argument is, of course, very familiar. When Head Start was originally proposed by President Johnson it was

intended to reverse the cycle of poverty and bring equity to school achievement, despite children's inequitable life circumstances (Zigler & Muenchow, 1992). At first many believed that this type of inoculation against the snares and traps of poverty could be overcome by just an eight-week summer Head Start program.

How will the NRS help us overcome the inequities of poverty? The answer implicit in the NRS is by demonstrating which programs are successful and which are not, so that poor performing programs can be improved or eliminated and high performing programs can be rewarded (Horn, 2003). Those who propound this accountability model are fond of likening the NRS or other accountability tests to a quality-assurance or quality-control system such as those used, for example, in manufacturing automobiles. Craig Ramey, chair of the Technical Work Group that advised Head Start about the test was quoted as saying, "If you were the head of any industry . . . you would have a quality assurance system in place to determine how your product is faring in term of quality . . . The Head Start test is just another quality assurance program" (Rimer, 2003, p. A23).

We know that although this country is very good at building factories, constructing assembly lines, and devising high-tech methods for producing goods, we have not been very successful at translating this expertise into our educational endeavors. As Malcolm Gladwell put it, "If schools were factories, America would have solved its education problem long ago" (Gladwell, 2003).

Why is the production model a poor fit for early education? One reason is that schools are not factories, children are not raw materials, and early care and education programs are anything but homogeneous stamping plants. The variables we work with are much more variegated and difficult to control than glass, steel, and production schedules.

Children enter preschool dramatically different one from the other. Just because nearly all of the children in Head Start are poor does not mean that they are all the same—and these differences go far beyond variation in geography, language, or ethnicity. Children also differ from one another in terms of inheritance, culture, experience, and many other factors. Moreover, as noted earlier, development is not linear, especially in the first five to eight years of life.

Simply put, the NRS was a failure not only because it tapped a narrow sample of children's skills; not only because a single purpose was never clearly specified; not only because it failed to conform to professional standards of test development; and not only because of its potential for changing Head Start to

a "skill and drill" curriculum. It was a failure because it ignored the complexity of early development that teaches us that no single indicator can assess a child's skills, achievements, or personality.

For these reasons and others, the House of Representatives passed an amendment to the Head Start Reauthorization bill in September 2005 to suspend the administration of the NRS until further information can be obtained about it from a panel of the National Academy of Sciences. Although the potential harm to children, teachers, and programs that could be done by the NRS will not be halted until the Senate approves a similar amendment, the NRS stands as a cautionary tale or paradigmatic illustration of the use of accountability testing in the early childhood years. No brief test of young children's achievement administered in a summative way can capture the complexity of pre–K children's growth. Just as a single facet of a reflective surface can never provide an accurate reflection of a complex phenomenon, so a one-dimensional early childhood test of achievement will give off a distorted image of what it is intended to measure. We will now turn to a proposal for evaluating children's growth and learning in early education programs that avoids the problems of the NRS and similar assessments.

No Easy Answers: Accountability and Evaluation

Wade Horn, the assistant secretary of HHS and one of the architects of the NRS, said the following about the test: "I can't for the life of me understand why anyone would think it's a bad idea to assess whether a program is progressing in crucial academic areas" (Friel, 2005, p. 541). Horn's comment is well-taken: there is nothing inherently wrong or mistaken about trying to find out how well a program is achieving its goals. Indeed, from a public policy perspective, it is critically important to have this information. The problem is that giving a test—especially, the NRS—appears not to be the best way to obtain this kind of information.

Measuring differences in children's vocabulary, letter knowledge, and early math skills does not give us information about program quality. The way to measure program quality is to gather data on such variables as low child-staff ratios; training of staff in early childhood development; provision of continuing professional development; use of practices that are developmentally appropriate; levels of positive interaction between children and staff; continuity, and competitive salaries and working conditions for staff; and the creation of a safe, caring

environment and one that encourages strong parental involvement (see, e.g., NICHD, 2000; Pianta, Howes, Burchinal, Clifford, Early, & Barbarin, 2005).

In order to learn about the impact of the kinds of variables listed above we cannot rely simply on a test of child outcomes or any other collection of unidimensional accountability data. Rather, what is needed is a design for a program evaluation, but one that does not preclude child outcomes.

Accountability calls for information about whether or not something happened. For example, was something taught or learned and how much was mastered? Evaluation data enable us to make inferences about why something happened. For example, why did this child learn more than that child, or why did this technique work better than that one? Accountability is close in concept to monitoring or documentation, whereas evaluation bears some resemblance to research into root causes of a phenomenon. Evaluation goes beyond the collection of child outcomes to include examination of variations in children, families, teachers, and programs that may help explain differences in those outcomes. In short, in program evaluation, the overall goal is to understand exactly what a program did and how it accomplished its purposes (Gilliam & Leiter, 2003).

In order to explain why a particular program works for a specific child under certain circumstances evaluation data must be collected on both structural and dynamic characteristics of the early childhood setting. Structural variables are those that represent formal features of early care and education programs that can be specified quantitatively and are often regulated by policy, such as class sizes and child/staff ratios. Dynamic factors are concerned with qualitative assessments of how teachers, staff, and children interact with one another on a regular basis and include an analysis of curriculum as well as teacher-child interactions. Finally, a whole range of demographic variables needs to be considered about the child and family as well as the teacher's background. These variables might include the child's age, gender, race/ethnicity, primary language, family socio-economic status, special needs, and previous out-of-home experiences. For the family we would need to explore socio-economic status, neighborhood characteristics, home environment, maternal mental health, and mother's age, education, and employment. We also want to know about teachers' age, gender, race, level of formal schooling, training in early childhood, amount and type of professional development, and teaching history. Without this information, we may fall into the trap of assuming that "one size" program fits all children, all parents, all teachers, and all communities.

However, for an evaluation design to be useful to policymakers we need to know more than just the structural and dynamic dimensions of a program. We also need to know the impact of the program on children. Our challenge is to accomplish this without incurring the numerous unintended consequences described earlier in this chapter.

One way of approaching this task is by constructing an assessment that is based on Item Response Theory (IRT). Such tests are intended to describe levels or patterns of growth, ability, or developmental achievement (Thorndike, 1999; Wright, 1999). They can be used as individually-administered assessments that provide information about a child's relative position on a specific developmental path ordered by difficulty.

As with virtually any other evaluative measure, IRT-based instruments are not immune to being used for high-stakes purposes. In order to minimize the potential for abuse and distortion, in this approach individual child scores would not be reported. Parents would receive an aggregated profile concerning the achievements of the children in their child's program. To learn more about their own child's accomplishments and areas in need of development, parents would meet with their child's teacher to obtain detailed information about the child's skills, accomplishments, and social-emotional characteristics. Information of this kind is available from instructional assessments that can provide diagnostic information about the child's learning. Such information can also be reported to and analyzed for policy makers (Meisels, Atkins-Burnett, Xue, Nicholson, Bickel, & Son, 2003).

Unlike other tests, IRT-based assessments can be administered without necessarily narrowing the curriculum. In developing IRT-based tests we try to determine whether different items represent estimates of the same level of achievement. In this way we can administer different sets of items that may all provide similar information to different children. In short, although it is essential that individual items are developmentally meaningful, they do not have the unique importance or status that is ascribed to items on other norm- or criterion-referenced tests. Consequently, they do not have the same potential to narrow the curriculum.

The importance of this is two-fold: (1) this approach may minimize teaching to the specific items of the test because they do not have the same value or meaning as on conventional tests and there are simply more of them in the test item bank than in a conventional test; and (2) we can assess change in children's

skills using a metric that describes a position on a developmental path, rather than strictly a position in a normative group. This enables us to focus our reporting on children's relative progress over time (this is known as a "value-added metric") rather than simply on achievement at the end of the program. To the extent that we are able to group children demographically according to comparability of background and to stratify our structural and dynamic program evaluation data according to program types and resources, it should be possible to draw conclusions about changes in developmental level or cognitive skills that are associated with the program itself.

Another important property of an assessment based on this model is that item difficulties can be tailored to the child's level of development, thus permitting more accurate measurement as well as minimizing the frustration levels for younger children and for children with special needs. This "tailoring" of item difficulty to the child's level of development minimizes floor and ceiling effects in cross-sectional, but more importantly, in longitudinal studies. Ideally, as exemplified by the Early Childhood Longitudinal Study-Kindergarten Cohort (West, Denton, & Germino-Hausken, 2000), we would use a two-stage, adaptive design in which children would first take a brief routing test to target the child's current level of functioning and assign them to a low, mid, or high level second-stage test. Then, on another occasion, they would be administered the second-stage test, which would permit a much more extensive review of the child's skills within that range of difficulty.

Although constructed differently from the model described above, several early childhood assessments based on IRT already exist (see Berry, Bridges, & Zaslow, 2004; and Kochanoff, 2003 for suggestions and reviews). Some of these instruments can be incorporated into evaluation designs, although the risk that their results will be misused and that they will become high-stakes assessments exists if they do not meet the other criteria presented earlier. After all, IRT was utilized in the development of the NRS, although many of the other features described here (e.g., large item bank, buffers against test-driven instruction, adaptive administration, sampling rather than census) were overlooked. No statistical technique, no matter how sophisticated, will solve all of the problems of high-stakes testing with young children. No statistical or psychometric technique is immune to misuse. Virtually all of the problems of classical norm-referenced testing described earlier can be ascribed to IRT as well, unless safe-

guards are put in place. But, if used as proposed, this plan will answer the policymakers' first question: Are children learning?

The second question—Are public funds being used wisely?—can also be answered by this approach because the information about structural and dynamic features of the programs will enable analyses to be conducted of a variety of program elements, pedagogical techniques, and child outcomes. In this way, it is possible to determine if particular aspects of the program, or the child and family background, are more or less strongly associated with child outcomes. This not only tells us if the program "works." It tells us for whom it works best under which combination of circumstances. For example, we may learn from a program evaluation that certain approaches to teaching reading are most successful with children from certain backgrounds. Evaluation data can also help us tailor specific programs for parents and particular inservice for teachers. The task here is to answer the overall question of program effectiveness by opening the "black box" of program operations and connecting these data to information about children, families, teachers, and children's achievements.

Conclusion

Accountability can have a meaningful role in early childhood if it is not monolithic in concept or high-stakes in implementation. Too often, policymakers and practitioners confuse means and ends when they discuss accountability. Tests and assessments are the means—or, more accurately, are among the means—that can be used to demonstrate accountability. But such measures are not sufficient to render a valid decision about whether a program is realizing its promise or achieving its goals. Assessments are administered for a wide variety of reasons, from screening to diagnosis and instructional planning to achievement testing. Not just any assessment makes sense in an accountability logarithm. And when high stakes are added to the equation, even more perturbations are generated. No summative test administered to individual children can tell us how well a program or school is performing overall unless we first know something about the background of the participants in that program (teachers, children, and families) and what the practices and processes of the program consist of. Added to this are the unique developmental features of early childhood, features that call into question the predictions that can be made from achievement tests early in life.

This paper has argued that high-stakes tests are of limited utility with young children and may even result in misleading conclusions and potentially damaging unintended consequences. In their place I recommend conducting a program evaluation on a sample of children and classrooms in order to provide a comprehensive picture of what children are learning; how they are being taught and by whom; and what the social context and resources of the program and families are. Evaluations based on a sample of children, teachers, and classrooms can provide us with useful and usable information that can answer our fundamental questions about children's learning and the value of our public investment in those programs.

Policymakers are notoriously unhappy with complex answers to apparently simple questions. Unfortunately, the questions discussed here—Are children learning? Are public funds being used wisely?—yield meaningful answers only when they reflect the complex phenomena they are intended to explicate. Simple answers that are flawed are of little value when compared with reliable data that are based on a comprehensive picture of what it is we want to know about. Although there are no easy answers to accountability in early childhood, the promise of embracing complexity is that more children will succeed because we will have the information we need to improve the programs that are providing them services.

References

Administration for Children and Families (June 6, 2003). *Information Memorandum: Description of the NRS Child Assessment*. Accessed online on September 28, 2003 at http://www.headstartinfo.or/publications/im03_07.htm

Berry, D. J., Bridges, L. J., & Zaslow, M. J. (2004). *Early childhood measures profiles*. Washington, DC: Child Trends.

Bowman, B. T., Donovan, M. S., & Burns, M. S. (Eds.) (2001). *Eager to learn: Educating our preschoolers*. Washington, DC: National Academy Press.

Brooks-Gunn, J., Fuligni, A. S., & Berlin, L. J. (2003) (Eds.). *Early child development in the 21st century: Profiles of current research initiatives*. New York: Teachers College Press.

Bronfenbrenner, U. (1974). *Is early intervention effective?* Washington, DC: Office of Human Development.

Bronfenbrenner, U. (1979). *The ecology of human development.* Cambridge: Harvard University Press.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion.* Norwood, NJ: Ablex Publishing.

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. In J. L. Herman & E. H. Haertel (Eds.). *Uses and misuses of data for educational accountability and improvement, 104th Yearbook of the National Society for the Study of Education* (Part II, pp. 289–319). Malden, MA: Blackwell.

Elmore, R. F. (2004). Conclusion: The problem of stakes in performance-based accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 274–296). New York: Teachers College Press.

Farran, D. C. (2000). Another decade of intervention for children who are low income or disabled: What do we know now? In J. P. Shonkoff & S. J. Meisels (Eds.) *Handbook of early childhood intervention.* (2d ed., pp. 510–548). New York: Cambridge University Press.

Friel, B. (February 19, 2005), Scrutiny mounts for Head Start. *National Journal,* 37(8), 539–541.

Goldstein, A., & Strauss, V. (July 8, 2003). Bush spells out Head Start changes. *Washington Post,* Page A02.

General Accountability Office (May 2005). *Head Start: Further Development Could Allow Results of New Test to Be Used for Decision Making.* Washington, D. C.: Author.

Gilliam, W. S., & Leiter, V. (2003). Evaluating early childhood programs: Improving quality and informing policy. *Zero to Three,* 23(6), 6–13.

Gladwell, M. (2003) Making the grade. *New Yorker,* accessed online on 9/30/05 at http://www.newyorker.com/printables/talk/030915ta_talk-gladwell

Halpern, R. (2000). Early intervention for low income children and families. In J.P. Shonkoff & S. J. Meisels (Eds.) *Handbook of early childhood intervention.* (2d ed., pp. 361–686). New York: Cambridge University Press.

Hart, K., & Schumacher, R. (2005). *Making the case: Improving Head Start teacher qualifications requires increased investment*. Washington, D. C.: Center for Law and Social Policy, Head Start Series, Paper No. 1.

Haskins, R., & Rouse, C. (2005). Closing achievement gaps. *The Future of Children, Policy Brief*. Washington, D. C.: Brookings Institution and Princeton University.

Herman, J. L. (2004). The effects of testing on instruction. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 141–166). New York: Teachers College Press.

Heubert, J. P., & Hauser, R. M. (Eds.) (1999). *High stakes: Testing for tracking, promotion, and graduation*. Committee on Appropriate Test Use. Washington, DC: National Academy Press.

Horn, W. F. (2003). Improving Head Start: A common cause. *Head Start Bulletin,* 76, 5–6.

Johnson, D. D., & Johnson, B. (2002). *High stakes: Children, testing, and failure in American schools*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Kagan, S. L. (1990). Readiness 2000: Rethinking rhetoric and responsibility. *Phi Delta Kappan,* 72, 272–279.

Kagan, S. L., Moore, E., & Bredekamp, S. (1995). *Reconsidering children's early development and learning: Toward common views and vocabulary.* Washington, D. C.: National Education Goals Panel.

Karoly, L. A., Greenwood, P. W., Everingham, S. S., Hoube, J., Kilburn, M. R., Rydell, C. P., Sanders, M., & Chiesa, J. (1998). *Investing in our children: What we know and don't know about the costs and benefits of early childhood interventions*. Santa Monica, CA: RAND.

Kim, J., & Suen, H. K. (2003). Predicting children's academic achievement from early assessment scores: A validity generalization study. *Early Childhood Research Quarterly,* 18, 547–566.

Kochanoff, A. T. (Ed.). (2003). *Report of the Temple University forum on preschool assessment: Recommendations for Head Start*. Philadelphia, PA: Temple University.

Kornhaber, M. L., & Orfield, G. (2001). High-stakes testing policies: Examining their assumptions and consequences. In G. Orfield & M. L. Kornhaber (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 1–18). New York: The Century Foundation Press.

LaParo, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years. A meta-analytic review. *Review of Educational Research,* 70(4), 443–484.

Lazar, I., & Darlington, R. (1982). Lasting effects of early education: A report from the Consortium for Longitudinal Studies. *Monographs of the Society for Research in Child Development,* 47, (2–3, Serial No. 195).

Lee, V. E., & Burkham, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin Kindergarten.* Washington, D. C.: Economic Policy Institute.

Lynn, R. L. (2005). *Fixing the NCLB accountability system*. CRESST Policy Brief 8. Los Angeles: UCLA Center for the Study of Evaluation.

Madaus, G. F. (1988). The influence of testing on the curriculum. In N. Tanner and K. J. Rehage (Eds.), *Critical issues in curriculum: Eighty-seventh yearbook of the national society for the study of education* (pp. 83–121). Chicago, IL: University of Chicago Press.

Madaus, G. F., & Clarke, M. (2001). The adverse impact of high-stakes testing on minority students: Evidence from one hundred years of test data. In G. Orfield & M. L. Kornhaber (Eds.). *Raising standards or raising barriers? Inequality and high-stakes testing in public education* (pp. 85–106). New York: The Century Foundation Press.

McNeil, L. M. (2002). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Mueller, J. (2002). Facing the unhappy day: Three aspects of the high stakes testing movement. *Kansas Journal of Law and Public Policy,* 11, 201–278.

Meisels, S. J. (1985). The efficacy of early intervention: Why are we still asking this question? *Topics in Early Childhood Special Education,* 5, 1–11.

Meisels, S. J. (1989). High stakes testing in kindergarten. *Educational Leadership,* 46, 16–22.

Meisels, S. J. (1992). Doing harm by doing good: Iatrogenic effects of early childhood enrollment and promotion policies. *Early Childhood Research Quarterly,* 7, 155–174.

Meisels, S. J. (1994). Designing meaningful measurements for early childhood. In B. L. Mallory & R. S. New (Eds.), *Diversity in early childhood education: A call for more inclusive theory, practice, and policy* (pp. 205–225). New York: Teachers College Press.

Meisels, S. J. (1999). Assessing readiness. In R. C. Pianta & M. M. Cox (Eds.), *The transition to kindergarten* (pp. 39–66). Baltimore, MD: Paul H. Brookes.

Meisels, S. J., & Atkins-Burnett, S. (2004). The Head Start National Reporting System: A critique. *Young Children,* 59 (1), 64–66.

Meisels, S. J., & Atkins-Burnett, S. (2006). Evaluating early childhood assessments: A differential analysis. In K. McCartney & D. Phillips (Eds.), *Handbook of Early Childhood Development* (pp. 533–549). Oxford: Blackwell Publishing.

Meisels, S. J., Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D. D., & Son, S. (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.

Meisels, S. J., & Shonkoff, J. P. (2000). Early childhood intervention: A continuing evolution. In J.P. Shonkoff & S. J. Meisels (Eds.) *Handbook of early childhood intervention.* (2d ed., pp. 3–33). New York: Cambridge University Press.

National Center for Early Education and Development (2005). Pre-K education in the states. *Early Developments,* 9(1).

National Institute of Child Health and Human Development Early Child Care Research Network (2000). Characteristics and quality of child care for toddlers and preschoolers. *Applied developmental science,* 4(3), 116–135.

National Institute for Early Education Research (NIEER). *The state of preschool: 2005 state preschool yearbook.* Rutgers University: Author.

Parini, J. (2005). *The art of teaching.* New York: Oxford.

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, D. C.: National Research Council.

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(3), 144–159.

Raudenbush, S. (2005). Newsmaker Interview: How NCLB testing can leave some schools behind. *Preschool matters,* 3(2), Rutgers University: National Institute of Early Education Research.

Rimer, S. (October 29, 2003). Now, standardized tests in Head Start. *New York Times,* p. A23.

Rist, R. C. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard educational review,* 40(3), 411–451.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom.* New York: Holt, Rinehart & Winston.

Ryan, J. E. (2004). The perverse incentives of the No Child Left Behind Act. *New York University Law Review,* 79(3), 932–989.

Schweinhart, L., & Weikart, D. (1980). *Young children grow up: The effects of the Perry preschool program on youths through age 15.* Ypsilanti, MI: Monographs of the High/Scope Educational Research Foundation, No. 7.

Shepard, L. A., & Smith, M. L. (1986). Synthesis of research on school readiness and kindergarten retention. *Educational Leadership,* 44, 78–86.

Shonkoff, J. P., & Phillips, D. A. (Eds.) (2000). *Neurons to neighborhoods: The science of early childhood development.* Committee on Integrating the Science of Early Childhood Development. Washington, D.C.: National Academies Press.

Stipek, D. (2005). Early childhood education at a crossroads. *Harvard education letter* [special issue], 21(4).

Thorndike, R. M. (1999). IRT and intelligence testing: Past, present, and future. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 17–36). Mahwah, NJ: Lawrence Erlbaum Associates.

West, J., Denton, K., & Germino-Hausken, E. (2000). *America's kindergartens: Findings from the Early Childhood Longitudinal Study, Kindergarten class of 1998–99, Fall 1998*. Washington, D. C.: US Department of Education, Office of Educational Research and Improvement.

White House Press Release/July 7, 2003. *President discusses strengthening Head Start*. Accessed online on December 12, 2005 at http://www.whitehouse.gov/news/releases/2003/07/20030707-2.html.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Lawrence Erlbaum Associates.

Zigler, E. E., & Muenchow, S. (1992). *Head Start: The inside story of America's most successful educational experiment*. New York: Basic Books.

Zigler, E. E., & Valentine, J. (Eds.). (1979). *Project Head Start: A legacy of the War on Poverty.* New York: The Free Press.

Herr Research Center for Children and Social Policy at Erikson Institute

The Herr Research Center for Children and Social Policy informs, supports, and encourages effective early childhood policy in the Great Lakes Region. The center generates original research and analysis that addresses unanswered questions about the optimal organization, funding, assessment, and replication of high-quality early childhood programs and services. Further, it provides comparisons of policies across states to determine which works best and why. Finally, through an array of publications, conferences, policy seminars, and advocacy efforts, it shares this research and analysis with state and local legislators, advocates, foundation officials, and other researchers in the field.

The center was established in 2005 with a gift from the Jeffrey Herr Family and grants from the Joyce and McCormick Tribune Foundations, as well as support from the Spencer Foundation and the Children's Initiative, a project of the Pritzker Family Foundation.

Center staff

Aisha Ray, Ph.D.
Acting Director

Eboni Howard, Ph.D.
Associate Research Scientist

Carol Horton, Ph.D.
Associate Research Scientist

Samuel J. Meisels, Ed.D.

Current research projects

After-school Programs
Analysis of Tuition-based Pre-K Programs
Chicago Program Evaluation Project
Early Childhood Mental Health
New Schools Project
New American Children and Families Project

Publications available from the Herr Research Center

Applied Research in Child Development

Number 1, After School Programs
Number 2, Father Care
Number 3, Welfare Reform
Number 4, Assessment
Number 5, Arts Integration
Number 6, Parent Support and Education

Occasional papers

"Lessons from Beyond the Service World," Judith S. Musick, Ph.D.
"Harder Than You Think: Determining What Works, for Whom, and Why in Early Childhood Interventions," Jon Korfmacher, Ph.D.
"Child Assessment at the Preprimary Level: Expert Opinion and State Trends," Carol Horton, Ph.D., and Barbara T. Bowman, M.A.
"'Does not.' 'Does too.' Thinking About Play in the Early Childhood Classroom," Joan Brooks McLane, Ph.D.
"Relationship-based Systems Change: Illinois' Model for Promoting Social-Emotional Development in Part C Early Intervention," Linda Gilkerson, Ph.D., and Carolyn Cochran Kopel, M.S.W.

Monographs

Critical Issues in After-School Programming, Robert Halpern, Ph.D.

Faculty

Samuel J. Meisels, Ed.D., President
Frances Stott, Ph.D., Dean/Vice President for Academic Affairs
Zachariah Boukydis, Ph.D.
Barbara T. Bowman, M.A.
Jie-Qi Chen, Ph.D.
Molly Fuller Collins, Ed.D.
Linda Gilkerson, Ph.D.
Robert Halpern, Ph.D.
Patty Horsch, Ph.D.
Jon Korfmacher, Ph.D.
Gillian Dowley McNamee, Ph.D.
Aisha Ray, Ph.D.
Sharon Syc, Ph.D.

Senior instructors

Collette Davison, Ph.D.
Mary Hynes-Berry, Ph.D.
Rebeca Itzkowich, Ph.D.

Senior research associates

Toby Herr, M.Ed.
Daniel Scheinfeld, Ph.D.

Senior research adviser

Charles Chang, M.A.

Erikson Institute

Herr Research Center for Children
and Social Policy

420 North Wabash Avenue
Chicago, Illinois 60611-5627

www.erikson.edu

Nonprofit Organization
U.S. Postage
PAID
Permit No. 2963
Chicago, Illinois