
Research Report


**Research Project 06-185**

# Evaluation Research in Education

October 10, 2007

BUREAU OF LEGISLATIVE RESEARCH

315 State Capitol Little Rock, Arkansas 72201 (501) 682-1937 www.arkleg.state.ar.us

# TABLE OF CONTENTS




## Acknowledgement

**Staff in the Bureau of Legislative Research wish to express their deep gratitude for the keen insights and valuable information provided by officials at the Arkansas Department of Education.**



**This report was produced by The Bureau of Legislative Research, Arkansas Legislative Council.**
**For further information or inquiries contact Dr. Brent Benda at Bendab@arkleg.state.ar.us or 682-1937.**

# Executive Summary

Among its mandates, the No Child Left Behind Act (NCLB) (2002) requires that funding for education and social programs be tied to empirical evaluation and analysis of effectiveness. Effectiveness is to be determined from *scientifically-based research.* A recent synopsis of state-of-the-art practices in education by well-established scholars concluded that evaluation research is a critical element of education reform in this country (Guthrie & Hill, 2007). These scholars state, "....there is no ready method for identifying...innovations, assessing their value, transmitting them to others, or combining several small ones into a broader innovation that might constitute a more productive way of teaching a whole course or grade level." (Guthrie & Hill, 2007, p. 6).

This report is presented in two separate sections: the first section presents a detailed discussion of the major principles and tools of program evaluation; the second section presents the existing program monitoring practices found in Arkansas, Kentucky, and Massachusetts. These particular states seem to have state-of-the-art practices for monitoring school improvement plans. However, no state was located that conducts statewide comprehensive program evaluations as described in Section 1 of this report. There are more limited evaluations of specific programs, but no state conducts a comprehensive evaluation of its educational interventions to determine their effectiveness in improving student learning, which ultimately is the goal of NCLB (Streifer & Schumann, 2005).

An intent of NCLB legislation is to require states to evaluate programs to determine what interventions effectively enhance achievement, for which students under what conditions. Developing effective and efficient educational programs requires systematic examination of the "effects" of programs on outcomes, such as student achievement, under varying circumstances. Comprehensive evaluations require examination of the influence of student, classroom, family, and community characteristics on the relationships between programmatic interventions and outcomes. As cogently argued by Guthrie and Hill (2007), accumulation of knowledge about effective education of students requires evaluation research by state departments of education. Systematic evaluation provides empirical confirmation or disconfirmation of more casual observations and professional judgments. The transparent reality is that various monitoring mechanisms, such as observational checklists and required reporting, do not provide data that has established reliability and validity, and too often no systematic effort is made to confirm putative relationships between inputs (interventions) and outcomes (Streifer & Schumann, 2005).

Presently, the states contacted for this report have developed rigorous auditing or review processes for program monitoring. This would appear to be a necessary first step to ensuring that school districts are complying with NCLB mandates. However, auditing and review procedures should not be confused with program evaluation procedures discussed in Section 1 of this report. The discussion is based on principles and practices presented in the professional literature on research methodology. Program evaluation essentially is research on intervention; it uses the same methodology as experimental or survey research. The most critical aspect of program evaluation is the examination of the "effects" of intervention on outcomes. The "gold standard" for evaluation is to have valid quantitative measures of both the intervention and outcomes to be able to examine the relationships between them. Rigorous evaluation requires valid measures of

both interventions (actions or programs) and outcomes, and systematic procedures (statistics) for examining the relationships between them (Streifer & Schumann, 2005).

In contrast, audit or monitoring procedures do not provide these evaluative measures. Except for quantitative outcome measures, such as norm-referenced testing, the data collected for auditing or review purposes are qualitative ratings by audit team members of performance on set standards. It is important to keep in mind that ratings rely on judgments. In Arkansas and Kentucky, these ratings must be based on three pieces of evidence (e.g., classroom observation, interviews with teachers). These requirements do provide some measure of assurance that idiosyncratic preferences are constrained in judgments. However, these procedures do not provide systematic evidence of the reliability and validity of ratings. Furthermore, no state systematically (statistically) examines the relationships between programs and outcomes (e.g., student achievement).

Comprehensive program evaluation involves three primary phases: 1) process evaluation, 2) outcome evaluation, and 3) impact evaluation. Process evaluation is requisite to an accurate interpretation of outcome and impact evaluations. Outcomes and impacts are a reflection of the quality of processes involved in implementing and maintaining programs. A process evaluation includes an extensive examination of the fit between theory and practical application, integrity of program implementation, training and qualifications of staff, linkages between services, clarity and measurability of goals and objectives, and reliability and validity of program and outcome measures.

In this report, the term *intervention* is used to include *programs* that are composed of services or actions in Arkansas Consolidated School Improvement Planning plans (ACSIP). *Services* or *actions* are elements (components) of a program that are used to intervene with clients to bring about desired changes (e.g., outcomes such as gains in student achievement).

Outcome and impact evaluations are concerned with the effectiveness of interventions (programs, services, actions) in achieving the changes stated in goals and objectives (e.g., gains in benchmark exam scores). Goals generally are more global statements that declare what programs will accomplish; whereas, objectives are measurable statements that specify: 1) who is targeted, 2) for what change, 3) how much change is expected, and 4) over what period of time.

The unit of analysis differentiates outcome from impact evaluation. Outcome evaluation examines changes in individuals' scores on exams (e.g., math or literacy exams) given before and after receiving an intervention (e.g., program, actions) for some designated period of time. Impact evaluation examines changes in aggregate scores (e.g., change in percentage of students who attained proficiency on math) and usually covers a longer period of time than outcome evaluation (e.g., changes over five years in percentage of students who are proficient). Outcome evaluation is concerned with the effectiveness of programs for individual students, whereas, impact evaluation is focused on effectiveness for an aggregate (e.g., students as a whole in Arkansas).

There are at least three primary ways that program evaluation differs from auditing or review practices being used to monitor school improvement plans. In program evaluation, a heavy

emphasis is placed on empirically establishing the validity and reliability of measures of interventions as well as outcomes and impacts. Also, comparisons are made between groups (treatment group versus control group) that receive and do not receive interventions to evaluate the effects of programs. When it is not practically or ethically possible to randomly assign students to these different groups, existing groups or a pre-test/post-test design is used. However, as discussed in Section 1, use of existing groups and pre/post-test designs are susceptible to attributing "effects" (changes in outcomes) to an intervention that are actually the result, fully or partially, of extraneous factors such as student or community characteristics. An optimal approach to dealing with extraneous factors in evaluation is to randomly assign students to an intervention group or to a control group.

Statistical procedures also have been developed to examine the separate *effects* (influences) of each extraneous factor, as well as the *effects* of actions and of the program as a whole. These statistical procedures permit a much more realistic analysis and evaluation of outcomes of interventions than simply observing gains in achievement and assuming that they are the result of antecedent actions or a program. The problem with this latter approach is that there is no objective evidence linking the intervention to the outcome or impact. The interpretation that an intervention (set of actions) led to achievement gains because it is antecedent to these gains assumes there are no extraneous factors. The assumption that there are no extraneous factors, however, belies a mountain of evidence showing that student learning is a multi-causal experience, and not the result of single experiences like educational intervention (Odden & Wallace, 2006).

This report presents a description of a hierarchical regression procedure that is specifically designed to examine complex relationships between extraneous factors, programmatic actions, and outcomes or impacts. Research evidence clarifies that student learning is the product of several ecological factors in addition to educational interventions, including individual characteristics, teaching, family dynamics, and school environment (Odden & Wallace, 2006). A comprehensive or accurate evaluation must take these factors into account.

In section 2 of this report, a description is provided of the scholastic audits used in Arkansas and Kentucky, as well as the two review processes employed in Massachusetts. As reported in that section, Arkansas has adopted the scholastic audit devised by Kentucky. Massachusetts was selected because it has an extensive review of schools and school districts, and it has the highest National Assessment of Educational Progress (NAEP) scores in the nation (http://nces.ed.gov/nationsreportcard/states/profile.asp). In fact, Massachusetts actually has two separate and distinct review processes, one conducted by the state department of education and one conducted by EQA, which is responsible to a citizen council appointed by the governor. The former review is of individual schools, whereas the latter is of school districts. Otherwise, the reviews in Massachusetts are similar to the audits done in Arkansas and Kentucky in using teams of experienced educators to rate schools on standards selected as important by the state department of education. In each review or audit, certain documents are required prior to an onsite visit by about six team members, who interview teachers, administrators, and parents. Until the current school year (2006-07), EQA emphasized managerial aspects of a district more than scholastic, whereas, the departmental review was more oriented to academic concerns. However, the EQA director reported that they are moving to an evaluation of curriculum and instruction to

determine what factors differentiate schools that are successful in raising student achievement scores from other schools.

Another difference among these three states seemed to be that students are interviewed in Arkansas and Kentucky, whereas, this did not seem to be true of Massachusetts. On the other hand, the EQA interviews municipal officials, which is not mentioned in the guidelines for audits in the other two states. Also, a random sample of successful schools in Kentucky (approximately 15%) and Massachusetts (40%) receive annual audits or reviews, respectively. The Arkansas Department of Education (ADE) began conducting scholastic audits in the 2006-07 school year, and audits are presently being conducted with schools that need improvement. Arkansas and Kentucky do have follow-up audits in which schools report on evidence regarding implementation and impact of improvement efforts.

The intent of this report is not to critique the auditing and review procedures implemented in these states because they serve a different purpose than program evaluation. Rather, the next logical step in school improvement, as suggested by the director of EQA, would seem to be to add program evaluation to auditing procedures to assess the effectiveness of programs being implemented. This report presents the key elements of program evaluation that will assist in the efforts to identify "what is working" in attaining student achievement.

In conclusion, existing audits and reviews of schools in state departments of education would seem to be providing very valuable diagnostic information and problem-solving recommendations to schools for planning improvements aimed at enhancing student achievement. The next logical step would be to evaluate these programs to confirm their effects on outcomes such as achievement scores or graduation rates.

# Section I
# Overview

Among its mandates, the NCLB Act requires that funding for education and social programs be tied to empirical evaluation and analysis of effectiveness. Effectiveness is to be determined from *scientifically-based research.*

The No Child Left Behind (NCLB) Act of 2001 (2002) was passed into law by the 107th Congress on January 8, 2002 (Public Law 107–110). This Act was intended to establish a new era of accountability for federally-supported education programs in the United States. Among its mandates, the NCLB Act requires that funding for education and social programs be tied to empirical evaluation and analysis of effectiveness. Effectiveness is to be determined from *scientifically-based research* (Mahoney & Zigler, 2006). According to NCLB, scientifically-based research is characterized by: 1) systematic empirical methods of observation and experiment; 2) rigorous and comprehensive data analyses that fully examine policy and practice issues; 3) measures and observational methods that provide valid data irrespective of evaluator or circumstance; and 4) acceptance from a peer-reviewed journal or a panel of independent experts, using comparatively rigorous, objective, and scientific review. (Public Law 107–110, 115 Stat. 1550–1551). The paramount importance and necessity of instituting rigorous research standards in evaluation of educational programs is clearly articulated in a recent scholarly review of the state-of-the-art in K-12 education (Guthrie & Hill, 2007).

Increasingly, policymakers and practitioners as well as researchers are using qualitative and quantitative methods of evaluation to identify best practices (Patton, 2002) in public health (Kalishman, 2006), in juvenile justice (Bradshaw & Roseborough, 2005), in environmental issues (Collier, 2006), and in education (Odden & Wallace, 2006). The overarching goal of evaluation research, or what is often referred to as program evaluation, is to contribute to the improvement of social conditions by providing scientifically credible information to decision-makers about the integrity and effectiveness of interventions.

## Purpose of Report

The purpose of this report is to discuss program evaluation in relation to existing practices of the Arkansas Department of Education.

The purpose of this report is to discuss program evaluation in relation to existing monitoring practices of state departments of education. This topic is relevant because of the increasing emphasis of NCLB on quantitative evaluation http://www.ncsl.org/programs/seminars/Forum /Index.htm).

It is not the intent of this report to critique existing monitoring practices because the states contacted appear to be taking the necessary initial steps mandated by NCLB. Furthermore, this report

is not intended to be a comprehensive review of existing practices, which would require a much lengthier document and investigative process.

This report specifies critical guidelines for process, outcome, and impact evaluations.

Instead, the report takes a more heuristic approach to discussing existing practices, and it discusses the major principles and tools of program evaluation. More specifically, this report specifies critical guidelines for process, outcome, and impact evaluations that need to be considered in the continued development of program evaluation.

These guidelines come from texts that several disciplines have considered to be landmark works for many years (Creswell, 2002; Patton, 2002; Rossi, Lipsey, & Freeman, 2004; Weiss, 1998).

## Definition and Purpose of Evaluation Research

Weiss (1998, p. 4) describes evaluation as "the *systematic assessment* of the *operation* and/or the *outcomes* of a program or policy, compared to a set of *explicit* or *implicit standards,* as a means of contributing to the *improvement* of the program or policy."

In an oft-cited book, Weiss (1998, p. 4; emphasis in original) describes evaluation as "the *systematic assessment* of the *operation* and/or the *outcomes* of a program or policy, compared to a set of *explicit* or *implicit standards,* as a means of contributing to the *improvement* of the program or policy." Weiss (1998, pp. 20–28) identifies several purposes for evaluating programs and policies as follows:

- Determining how clients are faring;
- Providing legitimacy for decisions;
- Fulfilling grant requirements;
- Making midcourse corrections in programs;
- Making decisions to continue or culminate programs;
- Testing new ideas;
- Choosing the best alternatives;
- Recording program history;
- Providing feedback to staff; and
- Highlighting goals.

Other reasons for conducting evaluations noted in the professional literature include: 1) accounting for how limited resources are used; 2) explaining what programs accomplish; 3) enhancing visibility of programs; 4) describing the impact of interventions; 5) increasing the efficiency of programmatic interventions; 6) supporting planning activities; and 7) providing evidence for decision making (Rossi et al., 2004).

In evaluation research, programs are composed of actions or services that are provided to targeted clients (e.g., students, teachers) in an intervention designed to bring about changes specified in goals and objectives. Goals are global statements about what will be accomplished by the programmatic intervention or actions.

Objectives state precisely: 1) what will change; 2) who is targeted for change; 3) how much change will occur; and 4) what is the time frame for change.

A few terms used in evaluation research need to be defined, since there is some variance in terminology among researchers. In evaluation research, programs are composed of actions or services that are provided to targeted clients (e.g., students, teachers) in an intervention designed to bring about changes specified in goals and objectives. Goals are global statements about what will be accomplished by the programmatic intervention or actions (e.g., according to NCLB, every child will be proficient in math, literacy, and science by 2014). These statements typically are not operationally definable or measurable, although this may vary from one evaluation project to another. Measurable or operationally defined statements generally are called objectives. Objectives state precisely: 1) what should be accomplished by each programmatic activity or service (achieve proficiency); 2) who is being targeted (e.g., all students in NCLB); 3) how much change (percent and degree) will occur (e.g., all students will achieve proficiency); and 4) in what time frame will the expected change occur (e.g., by 2014, all students will be proficient).

Operationally-defined objectives are statements that indicate clear markers (e.g., 10-point or 10% increase) or cutoffs (e.g., scores above 30 points) against which to evaluate change to determine if the objective has been achieved. These markers (or anchors) may indicate degrees of success (inadequate, minimal, adequate, sufficient, excellent), or they may be a single cutoff that distinguishes between effective and ineffective. This precision in statement of objectives may cause uneasiness if evaluation results are viewed as indictments or verdicts instead of systematic procedures for determining if interventions (actions, services) are effective in remedying problems for clients (e.g., students, teachers).

Furthermore, objectives are typically statements of relationship between an intervention and an outcome or impact (e.g., amount of change in student scores on a benchmark or normative test).

The following is a hypothetical example of an objective for a whole program:

> The XYZ Reading Program will elevate reading scores of males in grades 1-5 by 10 points during the 2006-07 school year.

Comprehensive program evaluations often contain objectives that pertain to more refined elements (activities or actions) that are components of that program. Objectives are specified whenever outcomes or impacts are desired from intervention efforts. Program evaluation should be conceptualized as a series of interlocking objectives whereby the outcomes of some objectives become the means of other objectives.

For example, the following objective may be requisite to the accomplishment of an objective which states that the school district will raise the literacy rate of its students by 10% over the next year:

> Purchasing materials to support the computer-based literacy program that will increase the use of computers by 50% among teachers during the 2006-07 school years.

However, the outcome of the objective for the action (purchasing materials to support the computer-based literacy program) is expected to contribute to the outcome stated in the objective for the XYZ Reading Program.

Manageability dictates that some combination of theory and professional judgment be used to select the actions that will be evaluated as objectives. Typically, only actions that are theorized to have a direct or indirect effect on outcomes are formally evaluated as objectives. As will be discussed, it is statistically possible to separate the effects of individual actions from the overall effects of the program on outcomes. Separating the individual effects permits the evaluator to determine which actions or combinations of actions are contributing to changes in the outcome. Separating effects of actions can be extremely important because it reveals which actions are effective in changing outcomes (e.g., increase in student scores) and which ones are basically superfluous or even detrimental. Schools can unknowingly take actions that work against their objectives, which would not be discovered without systematic analysis of individual effects of actions. Discontinuing actions that are superfluous or detrimental can be a big saving in terms of time and money (Odden & Wallace, 2006).

Systematic analyses of actions individually and in combination provide an understanding of what actions are actually contributing to the effectiveness of an intervention. This understanding can lead to savings in time and money because it can be generalized and applied in other schools districts.

Moreover, systematic analyses of actions individually and in combination provide an understanding of what actions are actually contributing to the effectiveness of an intervention. This understanding can lead to additional savings in time and money because it can be generalized and applied in other school districts. Instead of reinventing the wheel in every school district, information about which actions are effective could be shared across school districts. Currently, state departments of education require school

districts to justify action plans with evidence from the professional literature. Systematic statistical analyses of relationships between different actions and outcomes within states would provide more specific information about how effective specific programs or actions are in a particular state.

While the current assumption that effectiveness can be generalized across districts and states is likely true in several instances, there is considerable research that demonstrates that student and school characteristics affect the impact of programs on outcomes (Odden & Wallace, 2006). In fact, this is a primary reason NCLB requires states to disaggregate performance data by student characteristics. However, this evidence should not serve as an excuse for failure to implement programs. Programs may have to be tailored to be effective with certain students. Knowledge about how to tailor programs to meet the needs to particular students will only be developed by systematically evaluating intervention efforts.

As cogently argued in a recent critique of the K-12 education system, the imposition of a "one size fits all" approach to education programming in this country has stymied innovation and the systematic development of policies and procedures that have proven to be consistently effective in raising student achievement (Guthrie & Hill, 2007). Specifically, Guthrie and Hill (2007, p. 6) state, "Teachers and principals are constantly experimenting with new ideas in their schools and classrooms....[Yet] there is no ready method for identifying these innovations, assessing their value, transmitting them to others, or combining several small ones into a broader innovation that might constitute a more productive way of teaching a whole course or grade level."

Presently, states have focused on auditing programs for compliance with NCLB to the exclusion of evaluating these programs for their effects on outcomes such as student achievement and graduation rates. However, with little additional effort (time or resources), virtually the same monitoring procedures could be used to evaluate programs. The conversion of auditing procedures to evaluation is discussed in Section II of this report.

As clearly demonstrated by Guthrie and Hill (2007) in their overview of the state-of-the-art in education, evidence about what interventions are effective in enhancing student learning is seriously deficient because current knowledge-building mechanisms in this country are fragmented, isolated, and impeded by bureaucratic routines and political ideologies. Systematic program evaluation is essential to building knowledge about effective intervention

strategies (Streifer & Schumann, 2005).

## Types of Evaluation

*Program evaluations* typically proceed more or less sequentially through process, outcome, and impact phases.

Program evaluations typically proceed more or less sequentially through process, outcome, and impact phases (Patton, 2002; Rossi et al., 2004). Because each phase of program evaluation is sometimes performed to the exclusion of the others, these phases are often referred to as separate evaluations (e.g., process evaluation, outcome evaluation) -- a convention that will be used in this report. Process evaluation entails examining the processes of implementing, sustaining, monitoring, and altering services and programs. Qualitative and quantitative data are gathered on resources and training devoted to developing and maintaining a program, on the quality and integrity of program development, and on whether the intervention is reaching the entire targeted population. It should be noted at this juncture that some researchers use the terms *implementation* and *monitoring evaluation* instead of *process evaluation* for the same set of activities and procedures. *Implementation evaluation* is reserved for evaluations done in the embryonic stage of a program, whereas *monitoring evaluation* refers to evaluations that are done to assess the integrity or quality of services (actions) after the program has been running for awhile (Rossi et al., 2004). Process evaluation is concerned with services (actions) as well as programs (set of actions).

*Process evaluation* entails examining the processes of implementing, sustaining, monitoring, and altering services and programs.

Despite considerable agreement on evaluation methodology among researchers, there is no consensus on use of the terms *outcome* and *impact*. In fact, some evaluators use these terms interchangeably (Rossi et al., 2004). While, most researchers distinguish between *outcomes* and *impacts*, they do not agree on which term to assign to the two different levels of data analysis. In accord with NCLB mandates, the present report distinguishes between individual-level *outcomes* (e.g., changes in individual student achievement scores) and system-level *impacts* (e.g., statewide percentage changes in scores) in program evaluations. As the examples indicate, the same measures (i.e., student scores) may be used for both outcome and impact evaluations. When the same measures are used, it is the analyses and discussion of results that differentiate outcome from impact (Rossi et al., 2004). Frequently, however, outcomes and impacts are not based on the same data.

*Outcome evaluation* involves determining whether an intervention has achieved the desired changes among individual participants.

Outcome evaluation involves determining whether a program (intervention) has achieved the desired changes among individual participants in the program (Patton, 2002; Rossi et al., 2004). Objectives are written to state what outcomes are expected to result from the interventions (programs, services, actions) during a specified period of time. For example, have particular individuals' scores on math, literacy, or science exams increased a specified amount as a result of participation in a certain program over the past school year? The change in achievement scores over the past year is the outcome that is evaluated.

Whereas *outcome evaluation* addresses program effectiveness in terms of changes in individual scores, *impact evaluation* is concerned with the scope of change in a larger system (e.g., state).

Whereas outcome evaluation addresses program effectiveness in terms of changes in individual scores, impact evaluation is concerned with the scope of change in a larger system (e.g., school district or statewide system). For example, NCLB is concerned with the percentage of children who have achieved *proficiency* in math, literacy, and science in each year leading up to 2014, when all students are supposed to have achieved that level (Streifer & Schumann, 2005).

In sum, program evaluations are designed according to prescribed systematic principles and methods, which are more fully elaborated in some well-established books (Patton, 2002; Rossi et al., 2004; Weiss, 1998).

**Evaluation Designs**

The superiority of experimental methods for investigating the causal effects of deliberate intervention is widely acknowledged among researchers because of random assignment to treatment and control groups.

The classical methodological paradigm for program evaluation, especially of outcomes, is experimental design and its various quasi-experimental approximations (Creswell, 2002; Shadish, Cook, & Campbell, 2002). Experimental evaluation designs entail random assignment of persons to treatment (program) and to control (no treatment) groups, whereas quasi-experimental designs do not have random assignment to groups. The superiority of experimental methods for investigating the causal effects of deliberate intervention is widely acknowledged among researchers (Creswell, 2002). It is assumed that any factors that might influence outcomes and the impact of an intervention, aside from the program, get evenly distributed through random assignment of persons to the treatment (program) group or to the control group (i.e., those who do not receive the program).

At the same time, it is understood that the experimental paradigm is not pristine (Creswell, 2002; Shadish et al., 2002). While random

Random assignment to treatment and control groups is considered the ideal procedure in outcome and impact evaluations because it randomly distributes extraneous factors.

assignment to treatment and control groups is considered an ideal procedure in outcome and impact evaluations, decades of experience have shown important processes can occur after assignment that diminish the quality of the design, results, and utility of the evaluation. Among these processes are: 1) poor program implementation; 2) improvements by the control group unrelated to the intervention analyzed; 3) poor retention of participants in program and control conditions; 4) receipt of incomplete or inconsistent program services by participants; and 5) attrition or incomplete follow-up measurement. In addition, a host of participant characteristics (e.g. problem severity, motivation, ability) can interact with exposure and response to treatment in ways that further complicate evaluation (Shadish et al., 2002).

Extraneous factors are factors that influence outcomes and impacts in addition to interventions (e.g., student characteristics, family dynamics).

Known and measurable extraneous factors are analyzed with statistical procedures that can separate the effects of programs from the effects of extraneous factors.

The research term "extraneous factors" refers to program processes and contextual factors (e.g., biased sample, unique school characteristics) that influence outcomes and impacts in addition to the interventions. When extraneous factors are theoretically important and can be measured, they are incorporated in the evaluation and analyzed with statistical procedures that can separate the effects of programs from the effects of extraneous factors (see Rossi et al., 2004; and the discussion under the subheading Statistical Procedures for Evaluation). Stated differently, it is possible with these statistical procedures to examine the direct effects of intervention programs on outcomes, after considering (or controlling statistically) the influences of the extraneous factors. This approach not only permits statistical control over extraneous factors, but it also provides information about why particular outcomes are observed. For example, it might be found that the gains in student achievement in a school district are observed in only certain schools or only among specific groups of students.

Data collection has been increasingly extended in outcome evaluations to include measurement of such extraneous factors as program quality and client (e.g., student) characteristics.

Therefore, data collection has been increasingly extended in outcome evaluations to include measurement of such extraneous factors as program implementation (e.g., program quality and integrity), exposure to services (e.g., intensity, duration), client characteristics (e.g., economic disadvantage, gender), and responses that may mediate or moderate the effects of intervention. Mediation refers to factors diminishing or enhancing the effects of a program on an outcome, whereas moderation refers to modifying the programmatic effects in some way. For example, statistics might reveal that more years of teaching experience reduces (mediates) the effect of professional development training on assessed quality of teaching. Or, the results might show that teachers with bachelor's degrees derive more benefit from professional development training than those with master's degrees. In this latter case, the effects of

professional development on quality of teaching are moderated (modified) by academic degree. In other words, there is an interactive effect on teaching quality between academic degree and professional development training (Rossi et al., 2004). Collection of data on extraneous factors is a primary focus of process evaluation and is largely guided by theory.

Extraneous factors underlie the NCLB mandate that states disaggregate achievement scores by certain student characteristics.

Extraneous factors certainly are not novel ideas to educators. Indeed, extraneous factors underlie the NCLB (2002) mandate that states disaggregate student achievement scores according to economically disadvantaged students, students from all major racial and ethnic groups, students with disabilities, and students with limited English proficiency. More will be said about analyzing disaggregated data on outcomes in the section on Statistical Procedures for Evaluation.

**Use of Theory in Designing Evaluations**

Professional evaluators use a theory to identify important extraneous factors.

Because the potential number of extraneous factors can be almost limitless, professional evaluators use a theory or theories (or conceptual framework) to provide an explanation or rationale for why or how extraneous factors affect the outcomes and impact of programs (Rossi et al., 2004). Theory provides a roadmap (or model) for deciding what factors may mediate or moderate the effects of a program on an outcome or impact. When evaluation findings support or confirm theorized effects of extraneous factors and programs, professional evaluators are more confident in the validity of the findings because the theoretical model provides a rationale or explanation for the findings (Reynolds, 2005; Rossi et al., 2004). As Reynolds (2005, p. 2401) states, "Causal uncertainty is reduced through an examination of the empirical pattern of findings against the expectations inherent in the program."

Causal uncertainty is reduced through an examination of the empirical pattern of findings against the expectations inherent in the program theory.

As noted in a text on evaluation (Rossi et al., 2004), evaluators have long recognized the importance of a theoretical framework as a basis for formulating and prioritizing evaluation questions, designing evaluation research, and interpreting evaluation findings. *Program theory* has been given several names, including logic model, causal mapping, practice models, and action theory. Theory explains the relationships, however complex, between interventions, extraneous factors, and outcomes or the impact of programs (Patton, 2002; Reynolds, 2005; Rossi et al., 2004; Weiss, 1998). Stated succinctly, theory-based evaluation examines the pattern of interrelationships between all relevant influences within an intervention context (e.g., school district or state).

Theory explains the relationships between interventions (actions), extraneous factors, and outcomes or impacts.

The Consortium for Policy Research in Education at the University of Wisconsin-Madison has presented evidence that standards-based teacher evaluation systems constitute a performance competency model with the potential to improve instruction.

For example, the Consortium for Policy Research in Education at the University of Wisconsin-Madison has presented evidence that standards-based teacher evaluation systems constitute a performance competency model with the potential to improve instruction by affecting teacher selection and retention, motivating teachers to improve their skills, and promoting a shared conception of good teaching within school districts (Kimball, Milanowski, & Heneman, 2003; Kimball, White, Milanowski, & Borman, 2004; Milanowski & Kimball, 2005; Milanowski, Kimball, & Odden, 2005). These researchers theorized that standards-based teacher evaluation systems provide both incentives and guidance for teachers to change their teaching practices to conform to the standards embodied in the model presented. The standards model they presented is derived from the commonly used Foundations for Teaching authored by Danielson (1996a, 1996b; Danielson & McGreal, 2000). Danielson's standards for teaching are the basis for the rigorous performance assessment for experienced teachers used by the National Board for Professional Teaching Standards (http://www.nbpts.org/).

Mounting evidence confirms the theory that standards-based teacher evaluation systems enhance recruitment and retention of high quality teachers, lead to improved teaching skills, and encourage norms of excellence in instruction. (Odden, 2004; Odden, Borman, & Fermanich, 2004; Odden & Wallace, 2006). In terms of extraneous factors, Milanowski, Kimball, and White (2004) found that student characteristics (i.e., ethnicity, special education status, and English proficiency) did not mediate (e.g., reduce) the positive relationship between higher ratings of teaching and increases in student performance. In terms of moderation effects, they found that the positive relationship between quality of teaching and student performance weakens with teachers who have taught more than five years. In other words, more experienced teachers benefited less – albeit they did benefit – from standards-based evaluation than their inexperienced counterparts. These moderation effects support the theory that standards-based teacher evaluation systems encourage the recruitment and retention of high quality teachers. According to the theory, less skilled teachers tend to shun or leave schools districts with standards-based evaluation because they are not motivated to excel, or they recognize their limited abilities to meet expectations (Odden & Wallace, 2006).

Programs and evaluations that are not based on explicit theories rest on implicit theories often characterized as intuitive hunches, common sense, practice experience, or professional wisdom. The point is that every intervention, whether or not it is consciously acknowledged, is based on theory.

Programs and evaluations that are not based on explicit theories rest on implicit assumptions often characterized as intuitive hunches, common sense, practice experience, or professional wisdom. The point is that every intervention, whether or not it is consciously acknowledged, is based on theory (Guba, 1990). A problem with unknown or implicit theories is that they rarely acknowledge key extraneous factors. Yet, evidence is clear that extraneous factors play a major role in the context of educational programming (Odden et al., 2004).

The ADE currently does not conduct evaluations of programs within the guidelines being discussed in this report. For example, they do not present a formal conceptual model for evaluating programs. They do not statistically examine relationships between extraneous factors, interventions, and outcomes or impacts. Instead, assumptions are made that any gains in student achievement are the result of actions taken preceding the observed changes.

Current scholastic audits of school programs used in the states studied are not conceptualized with theories about relationships between extraneous factors, interventions, and outcomes or impacts. While they do disaggregate data on student achievement according to NCLB (2002) mandates, they do not formally (statistically) analyze these data in terms of how actions taken in programs and extraneous factors affect outcomes. In fact, there does not seem to be a systematic effort to verify that actions or programs have certain outcomes. Rather the current approach seems consistent with the label "scholastic audit," where emphasis is placed on making sure actions taken are consistent with improvement plans presented to the state department of education. There appears to be no emphasis on measuring program elements and outcomes and systematically analyzing relationships between them. Instead, assumptions are made that any gains in student achievement are the result of actions taken preceding the observed changes. While these assumptions are intuitively plausible, they are not based on any direct evidence to indicate they are valid.

There are several potential problems methodologically with assuming antecedent actions necessarily resulted in any subsequent changes. Foremost, there are several possible extraneous factors that may have indirectly or directly affected the gain in student scores, especially when any amount of gain qualifies as improvement. A particularly salient extraneous factor, according to research (Odden et al., 2004), which is not considered in existing reviews, is quality of teaching (Odden & Wallace, 2006). Furthermore, it is plausible that teachers increase their efforts, or focus their attention on certain topics, when their school is placed on the "needs improvement" list. In other words, any gains in student performance may be the product of quality and focus of teaching instead of actions presented in school improvement plans.

Furthermore, because existing audit procedures in state departments of education do not entail any statistical analyses of relationships between actions (intervention) taken by schools and outcomes (e.g.,

changes in student learning scores), there is no information on which actions or combinations of actions may contribute to any outcome gains observed. This approach means that there is no accumulation of objective (or verified) knowledge about what interventions are effective in bringing about desired outcomes.

## Phases of Program Evaluation

### Process Evaluation

Process evaluation is the systematic examination of services (program activities or actions) and processes (procedures of implementation and administration of programs) to determine how well they are operating and how well they conform to the plans and expectations for the program (Rossi et al. 2004). Included in a process evaluation is:

- extensive examination of the fit between theory and practical application;
- integrity of program implementation;
- training and qualifications of staff;
- linkages between services;
- clarity and measurability of goals and objectives;
- availability and adequacy of resources;
- quality and integrity of intervention;
- clear specification of outcomes and impacts;
- reliability and validity of measures of program processes, outcomes, and impact;
- program quality and intensity; and
- sufficiency to meet needs.

Process evaluations include qualitative and quantitative data from several sources. One aspect of process evaluation, for example, might entail in-depth individual interviews or a focus group discussion with staff and administration to learn as much as possible about planning, designing, implementing, monitoring, and altering various services and processes that compose programs (Patton, 2002). These qualitative data often provide valuable insights into decision-making and operations that contributed to success or failure of services and processes. These interviews and focus groups yield more valuable information when they are conducted by experienced facilitators who use a combination of systematic and open-ended questions and prompts (Patton, 2002).

The more qualitative data can be confirmed or disconfirmed by using client- and staff-satisfaction surveys that address similar content. Satisfaction surveys are very valuable as an aspect of process evaluation, but they should not be the sole basis for evaluation. True evaluation requires some quantitative data as well as analyses of outcomes and the impact of programs. In fact, most research evaluators desire to triangulate different forms and sources of data (Patton, 2002; Rossi et al., 2004). For example, the quality of teaching has been evaluated by triangulating traditional unrestrictive observations by an administrator with systematic ratings from standards-based assessments conducted by multiple raters (e.g., Milanowski et al., 2005; Odden et al., 2004).

Process evaluations can be conducted throughout the existence of a program to determine: 1) size and appropriateness of clientele; 2) amount, type, and quality of services; 3) if eligible subgroups are served; 4) if all eligible subgroups are informed about the program; 5) if staff is sufficient in number, training, and skills; 6) if services and procedures are well coordinated; 7) extent of collaboration between program and other agencies; 8) if there are adequate facilities and funding; 9) compliance with various standards; 10) staff and client satisfaction; and 11) type and extent of follow-up.

It is important to recognize that many of the questions addressed by process evaluations require judgments about levels of performance (e.g., appropriate, satisfactory, reasonable, adequate). Judgments are based on implicit or explicit criteria. There are several approaches to setting criteria for program performance, including theory, values, and statistics (http://www.anaesthetist.com/mnm/stats/roc/Findex.htm).

The most common and widely accepted criteria for making judgments in practice are administrative standards, which are typically based on practice experience, research findings reported in the professional literature, and consensus among policymakers (Patton, 2002; Rossi et al., 2004). Some aspects of program performance may fall under applicable legal, ethical, or professional standards, such as NCLB mandates.

Moreover, the assessment of particular dimensions of program performance often is not based on predetermined criteria, but represents an after-the-fact judgment call. What is important to recognize is that judgments are based on some criteria. These criteria should be made explicit and carefully examined in light of the mission and goals of the system, as well as existing evidence.

As alluded to earlier, program process is often the focus of formative evaluations designed to provide useful feedback about how well the program is implemented, administered, and monitored, and how well it reaches the targeted population. Another major role of process or implementation evaluation is to provide a context in which to properly interpret outcome and impact evaluations (Rossi et al., 2004). Indeed, outcome and impact evaluations, without a process evaluation, can provide erroneous data when programs are poorly implemented and maintained. For instance, the lack of student gains in achievement may be due to a poorly run program rather than to the program per se.

Another phase of process evaluation is monitoring, which provides information on how a program is performing its critical functions after it has existed for awhile. This type of feedback allows managers to take corrective action when problems arise and it provides periodic assessment of program performance. Ideally, the monitoring activities undertaken as part of evaluation should meet the information needs of all constituencies (administration, policymakers, sponsors, and stakeholders) (Rossi et al., 2004).

**Outcome Evaluation**

"Assessing a program's effects on clients it serves and the social conditions it aims to improve is the most critical evaluation task because it deals with the 'bottom line' issue for social programs" (Rossi et al., 2004, p. 204).

A widely influential book in research evaluation (Rossi et al., 2004) states:

> "Assessing a program's effects on clients it serves and the social conditions it aims to improve is the most critical evaluation task because it deals with the 'bottom line' issue for social programs. No matter how well a program addresses target needs, embodies a good plan for attack, reaches its target population and delivers apparently appropriate services, it cannot be judged successful unless it actually brings about some measure to beneficial change in its given social arena." (p. 204).

In other words, Rossi et al. are clearly stating that while process evaluation is necessary, it is not sufficient for comprehensive program evaluation. This observation is important because existing state audits and reviews are composed of practices that parallel process evaluations, and do not include the critical element of outcome evaluation. Rossi et al. (2004, p. 204) define an outcome as the state of the target population or the social conditions that a program is expected to have changed. They note that outcomes are observed characteristics of the target population (e.g., students, parents) or social condition (e.g., low literacy in the state), not of the program. Outcomes represent the level of benefit that clients

receive from services (actions, programs), not simply the receipt of services. In education, for instance, it is the gains in student achievement that are the desired outcome of programs, rather than whether certain students received the program. Whether or not all targeted students receive the program is a concern of process evaluation. (Rossi, et al., 2004, p. 206).

Comprehensive outcome evaluations are based on explicitly articulated theory that links programs or services (actions) to intermediate outcomes that, in turn, are expected to lead to more long-range outcomes. If correctly stated, this series of linked relationships among factors (actions and outcomes) represents the assumptions about the critical steps between program services (actions) and the ultimate intended benefits of the program. Theory is a set of logical assumptions (statements) that explain the change in the outcome desired. Because outcomes are the primary purpose of evaluation, they should be observed with reliable and valid measures.

**Reliable and Valid Measures in Evaluation**

The reliability of a measure (e.g., math test) is the extent to which the measure produces the same results each time it is administered.

The reliability of a measure (e.g., math test) is the extent to which the measure produces the same results each time it is administered. Identification and measurement of outcomes, to the extent possible, should be informed by theory and evidence found in the professional literature. Comprehensive theory describes the properties or elements that comprise concepts (e.g., student learning), and often methods (e.g., scales) are developed directly from theory to observe (measure) those properties (Corcoran & Fischer, 2000). Measurement scales are constructed of items that exemplify elements of concepts. The effect of unreliable measures is to have unreliable results.

Generally, evaluators try to locate measures that have been established as reliable and valid in prior research. However, there are times when existing measures are not available or suitable. In this case, there are several psychometric methods for determining the reliability of a scale.

Generally, evaluators try to locate measures that have been established as reliable and valid in prior research (Rossi et al., 2004). However, there are times when existing measures (e.g., scales, instruments, questionnaires) are not available or suitable (e.g., too time-consuming). In this case, there are several psychometric methods for determining the reliability of a scale (or set of items) created for a specific project (Carmines & Zeller, 1979; Devellis, 2003; Netemeyer, Bearden, & Sharma, 2003). Reliability can vary according to sample and circumstances of measurement, so it cannot be automatically assumed that an established measure will perform in the same way in a different sample or set of circumstances (Rossi et al., 2004). Therefore, consideration has to be given to the characteristics of samples and circumstances that

have been used to determine reliability of a measure. Too, most research evaluators examine the psychometric properties of measures in their sample.

The validity of a measure is the extent to which it measures properties of a concept it is intended to measure.

The validity of a measure is the extent to which it measures properties of a concept it is intended to measure. The question that validity answers is whether, for example, an IQ test actually measures intelligence, or a math exam measures math abilities.

Related to validity, evaluators are especially interested in the *sensitivity* of measures to change because program evaluations are primarily aimed at assessing change in an outcome (e.g., change in students' scores). Measures can be insensitive to change because they do not actually measure the particular skill targeted by a program (e.g., a generic test may not measure math problem-solving). Also, some measures (self-esteem) are not designed to indicate changes induced by relatively brief interventions. The point is that measures must be selected carefully so outcomes and impacts can be accurately measured and interpreted (Rossi et al., 2004).

There are books available that contain measures and descriptions of psychometric properties of those measures, such as reliability, validity, sensitivity, and circumstances under which these psychometric properties were established (Corcoran & Fischer, 2000).

**Impact Evaluation**

What differentiates impact evaluation from outcome evaluation is the unit of analysis. That is, whereas outcome evaluation may focus on changes in individual scores on math or literacy exams over a school year, impact evaluation would be centered on changes in some aggregate scores (e.g., change in percentage of students who score proficient in the state or a school district). Furthermore, impact evaluation typically is concerned with a longer period of time than outcome evaluations, such as the change in percentage of students who score proficient between now and in 2014. The unit of analysis is individuals in outcome evaluation, while it is an aggregate in impact evaluation.

The unit of analysis is individuals in outcome evaluation, while it is an aggregate in impact evaluation.

This discussion of the reliability and validity of measures and extraneous factors emphasizes the fact that findings (e.g., gains in student achievement) do not speak for themselves as commonly asserted. Findings on outcomes are a product of programs, measurement, and extraneous factors. The interpretation of any set of findings must consider whether the findings are based on solid

measures and on complete analysis of all pertinent influences on the outcomes (Patton, 2002; Rossi et al., 2004).

**Design of Impact and Outcome Evaluation**

All evaluations of programs are inherently comparative. Determining the outcome or impact of a program requires comparing the condition (e.g., student math scores) of participants that have experienced an intervention with an estimate of what their condition would have been had they not received the intervention (Rossi et al., 2004, p. 236). In practice, this comparison often involves a comparison of outcomes between program participants and an equivalent group of persons who do not receive the program (i.e., control group).

There are several approaches to establishing an equivalent or control group for comparison with the so-called program (or treatment) group (persons who receive the intervention). A popular approach is a pre-test/post-test design whereby a comparison is made between the scores individuals make before (control group) and after an intervention (program group). While this design is intuitively compelling because the intervention appears to be the most logical *cause* of any changes in scores, it is not as convincing to researchers as a randomized field experiment (Rossi et al., 2004, p. 237). That is, researchers understand that it is impossible to conduct a perfect evaluation. All evaluations are limited by imperfect methodology, including imprecise measures, non-representative samples, and procedural problems. Researchers refer to these methodological limitations as biases (Patton, 2002; Rossi et al., 2004; Shadish et al., 2002).

All evaluations are limited by imperfect methodology, including imprecise measures, non-representative samples, and procedural problems. Researchers refer to these methodological limitations as biases.

A huge advantage of the randomized field experiment over the pre-test/post-test design is that biases are randomly assigned to both the treatment and control groups. By contrast, biases often remain undetected and unknown in the pre-test/post-test design. A particularly troublesome problem with the pre-test/post-test design is growth bias (Rossi et al., 2004, pp. 268-269). Consider, for example, a literacy program for young children that emphasizes vocabulary development. Let us assume we have a reliable, valid, and sensitive outcome measure of vocabulary, and that we use this measure to get individual scores on vocabulary before and after a year-long tutoring program. In using this pre-test/post-test design, we are assuming that the children's vocabulary would remain virtually the same without the tutoring when we report the gain in scores observed during the year evaluated. However, studies of child development disprove this assumption of no substantive

change in vocabulary without intervention (e.g., Aarnoutse, van Leeuwe, & Verhoeven, 2005). Moreover, developmental changes (growth bias) are not the only extraneous factors that may influence (or bias) the outcome or impact. Illustratively, it is plausible that teachers might put special emphasis on vocabulary in their classes after their school district institutes a tutoring program and exams aimed at vocabulary. Hence, gains in vocabulary attributed to tutoring in fact may be largely due to a combination of natural developmental changes and special emphasis on vocabulary by classroom teachers.

The pre-test/post-test design does not permit separation of the effects of actions from the effects of extraneous factors. Therefore, gains attributed to particular actions may be the result of a combination of unmeasured extraneous factors.

A simple pre-test/post-test design does not permit separation of the effects of the program (actions) from the effects of extraneous factors, such as developmental changes and classroom emphasis. Therefore, gains attributed to particular actions may be the result of a combination of unmeasured extraneous factors.

To be certain that outcomes and the impact observed are a result of programs, extraneous factors can be measured and analyzed together with program effects to determine the relative contribution of each influence on the results.

There are three major methodological remedies used by evaluators to overcome biases introduced by extraneous factors: 1) measuring and analyzing known extraneous factors; 2) adding a control group to the design; and 3) randomly assigning students to program and control groups (Rossi et al., 2004). Each remedy independently adds strength to the design of the evaluation, and together they are considered the gold standard for designing evaluation (Rossi et al., 2004, p. 237). As Rossi et al. (p. 269) note, the beauty of random assignment of individuals to program and control groups is that we can make the assumption that the groups are equivalent within the bounds of chance fluctuations that can be assessed with statistical tests. Any extraneous factors (biases) get randomly distributed to the groups. To be certain that outcomes and the impact observed are a result of programs, extraneous factors can be measured and analyzed together with program effects to determine the relative contribution of each influence on the results. In other words, the effects of the program can be separated from the effects of various extraneous factors (see section on Statistical Procedures for Evaluation on page 31 of this report).

There are actually two ways in which program effects are observed and accounted for in evaluation analyses. A common approach is to simply compare students who have received some intervention to those who have not received it. An even more compelling approach

Observed statistical relationships between measures of a program and outcomes or impact provide an even stronger empirical case for the effectiveness of programs than simply observing statistical differences between treatment and control groups.

is to measure the aspects (e.g., quality, intensity, frequency, duration) of a program that are theorized to result in the outcomes or impact. Observed statistical relationships between measures of a program and outcomes or impact provide an even stronger empirical case for the effectiveness of programs than simply observing statistical differences between treatment and control groups (Rossi et al., 2004).

Random assignment to groups gets rid of the *selection bias* introduced by the more frequently used nonequivalent comparison design.

Random assignment to groups gets rid of the *selection bias* introduced by the more frequently used nonequivalent comparison design, which involves systematically assigning people to treatment and control groups. For example, two schools are often compared with the assumption that the one receiving the program is equivalent to the one used as a control group. However, without convincing measures of several relevant extraneous factors, the assumption of equivalence of groups is at best tenuous. A nonequivalent comparison design is always vulnerable to post hoc discoveries of unforeseen extraneous factors that differentially influenced the program and control groups. There are many personal, cultural, and economic factors that influence residence, school attendance, and exam scores.

Most researchers are not willing to assume groups are equivalent without convincing evidence. Therefore, while adding a control group can give some indication of growth (e.g., gain in math scores result from natural reasoning and socialization) without intervention, it can introduce the problem known as *selection bias.* Selection bias refers to extraneous influences that are introduced into evaluation analyses by non-random or systematic selection (e.g., administrative assignment of students to schools) into program and control groups. If these extraneous influences are not measured and analyzed, there is no way to know how much influence they have on outcomes or impacts. As a consequence, changes in an outcome or impact attributed to programs may actually be the product of several unknown extraneous factors (e.g., economic disadvantage, quality of teaching, differences in school resources).

The problem of *selection bias* is often compounded by differential attrition. A comparison of two schools, for instance, can end up being biased because of differential dropout or migration rates between the schools.

Another form of attrition that can degrade all research designs is lack of motivation to provide valid responses. This form of attrition generally is more likely among members (e.g., teachers and students) of the control group because they often lack the motivation

to fully participate that is instilled by offering services to people.

## Controlling for Biases in Design

While random assignment to program and control groups is the optimal evaluation design, it is not always ethically, politically, or practically feasible.

While random assignment to program and control groups is the optimal evaluation design, it is not always ethically, politically, or practically feasible (Rossi et al., 2004). Reality often dictates that a quasi-experimental design be used; that is, people are systematically assigned to a program group and to a control group. One way of attempting to ensure equivalence of these groups is to match individuals or an aggregate (e.g., schools). This approach, however, assumes that the relevant extraneous factors are known and can be matched. As Rossi et al. (2004, p. 279) observe, "However carefully matching is done, there is always the possibility that some critical difference remains between the intervention group and the selected controls." It is also true that matching is very difficult at best (Rossi et al., 2004).

The commonly-used approach to achieving equivalence of groups in recent years is the use of statistical controls.

Therefore, the commonly-used approach to achieving equivalence of groups in recent years is the use of statistical controls (Rossi et al., 2004). There are several multivariate statistical procedures that can separate the individual effects of extraneous factors from the effects of the program. These procedures account for (or statistically control) the initial differences (on extraneous factors) between the program and control groups by subtracting out the portion of variance in outcomes (e.g., student gains in math scores) attributable to these initial differences from the portion that is the result of the program.

## Statistical Procedures for Evaluation

The particular statistical procedure selected to control extraneous factors depends on characteristics of the measures (e.g., level of measurement), form of theorized relationships between factors, the statistical assumptions deemed realistic, and the knowledge of the evaluator (Rossi et al., 2004). Most commonly, some variant of multiple regression procedures (Freund & Wilson, 1998) or structural equation modeling (Bollen, 1989; Maruyama, 1998) is used to determine the separate effects of extraneous factors and the intervention program. The value of any statistical approach is largely determined by the inclusion of relevant extraneous factors in the analysis of the effects of a program on an outcome or impact. If an extraneous factor that has a significant influence on an outcome is left out of the analysis, then any findings regarding the effects of a program are distorted by that unexamined extraneous influence. The more extraneous influences that are not examined in an evaluation

analysis, the more distortion one gets in examining the relationship between an intervention and an outcome. As stated earlier, an extraneous influence can distort the *effects* of the program by either mediating (strengthening or diminishing) or moderating (modifying) these effects. Multivariate regression procedures and structural equation modeling can identify and determine the amount of influence of each mediating, moderating, and intervention effect on an outcome or impact (Rossi et al., 2004).

The evaluator has to identify and measure extraneous factors that are likely to mediate or moderate the effects of a program on an outcome or impact. Influential extraneous factors typically are identified in theories and in the professional literature, and they are often known in the practice arena (Patton, 2002; Rossi et al., 2004). For example, NCLB (2002) requires states to disaggregate data according to economically disadvantaged students, students from all major racial and ethnic groups, students with disabilities, and students with limited English proficiency because research has indicated that these extraneous factors distort the effects of educational programs on student achievement gains. For example, when achievement scores are statistically controlled (held constant) for economically disadvantaged students, it might be observed that there are greater gains in overall achievement in a state. If the information about who is economically disadvantaged had not been collected, and their achievement scores had not been statistically controlled, then analyses would have indicated that the program had less affect on student achievement. In statistical terms, being economically disadvantaged mediates – in this case, lessens – the effects of the program on student achievement gains.

While student scores can be disaggregated by extraneous factors and analyses done in separate aggregates (e.g., economically disadvantaged), this approach does not reveal the relative effects of extraneous factors and a program. Moreover, disaggregation can quickly reduce the number of scores analyzed (e.g., students with disabilities) to a meaningless level, especially in small school districts. By contrast, multivariate regression procedures can simultaneously consider (in one equation or analysis) the effects of several extraneous factors on student achievement and several actions that might comprise a program (or an intervention) (Reichardt & Borman, 1994).

Multivariate regression procedures can simultaneously consider the effects of several extraneous factors on student achievement and several actions that might comprise a program.

Stated succinctly, multiple regression procedures (e.g., ordinary least squares, logistic) provide precise information about how much influence each factor (extraneous and programmatic actions or services) has on an outcome or impact measure (e.g., student

Multiple regression procedures provide precise information about how much influence each extraneous and programmatic factor has on an outcome or impact measure.

achievement gains). There are three primary coefficients derived from regression analyses that indicate how much influence factors are having on an outcome or impact: 1) unstandardized regression coefficient (β); 2) standardized regression coefficient (Beta); and 3) amount of variance in the outcome (or impact) accounted for by the predictors ($r^2$). Factors (extraneous and program actions) are referred to as predictors in regression analyses. β indicates amount of change in the outcome (e.g., change in student scores) with every unit change in a predictor (e.g., program action). For example, a β of 12.45 for a computer-based literacy program would indicate that this action or program increased student scores in literacy by twelve and a half points. In binary data, the unit change in predictor indicates the comparison of being in one category (e.g., program group) instead of the other category (e.g., control group). β tells you precisely how much change in scores can be attributed to each predictor. Beta, in contrast, indicates the relative effects of each predictor. Typically, the beta coefficient varies between 0 and 1, with 0 indicating that a factor does not predict the outcome (e.g., student scores), whereas 1 indicates perfect prediction. For example, if student achievement gains are predicted by gender (beta = .10), economic status (beta = .20), and a computer-based literacy program (beta = .40), these results would indicate that the program used is four times more predictive of student gains than gender and twice as predictive as economic status.

Finally, $r^2$ indicates how much of the variance in an outcome (e.g., gains in student achievement scores) is accounted for (or explained by) by each predictor. $r^2$ varies between 0 and 1, with 1 indicating that all variance is accounted for by predictors, whereas 0 shows that no variation is explained.

There are several approaches to multiple regression statistics used by evaluators.

There are several approaches to multiple regression statistics used by evaluators (Rossi et al., 2004). For example, there is a one-stage regression procedure where all predictors of an outcome are considered simultaneously in a single equation. This procedure may examine factors that are related to selection into a program or control group, but it does not analyze these factors as predictors of that selection. An alternate approach that is becoming more commonplace is a two-stage procedure in which the first step is to use relevant extraneous factors to construct a statistical model that predicts selection into the program or control groups. The second step is to use the results of the first step to combine all the extraneous factors into a single composite selection variable. The selection variable is then used as a control factor in the analysis of the effects of a program on an outcome. This two-stage procedure in which the first stage attempts to statistically describe the

differential selection of individuals into nonrandomized program and control groups is called *selection modeling* (Rossi et al., 2004, p. 285). Several variants on selection modeling are available, including Heckman's (Heckman & Hotz, 1989) econometric approach, Rosenbaum and Rubin's (1984) propensity scores, and instrumental variables (Greene, 1993).

Multi-level modeling is especially valuable in evaluation because it does account for the larger context that contains important influences on children's learning, including individual characteristics, teaching, classroom and school environments, and familial dynamics.

A particularly sophisticated regression procedure that considers the fuller ecological context (e.g., student, teacher, classroom, school district) of student learning is multi-level (hierarchical) modeling – also known as hierarchical linear models (Gelman, 2005; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Multi-level modeling is especially valuable in evaluation because it does account for the larger context that contains important influences on children's learning, including, but not limited to: 1) individual characteristics (e.g., intelligence, learning disabilities); 2) teaching, classroom and school environments; and 3) familial dynamics (Gelman, 2005). Succinctly stated, multi-level modeling is valuable to evaluation because it more accurately reflects the reality of student learning than the more truncated statistical approaches that only consider select aspects of learning, such as the effect of an intervention on some outcome (see Kimball et al., 2003; Kimball et al., 2004; Milanowski & Kimball, 2005; Milanowski et al., 2005; Odden & Wallace, 2006).

## Section 2
## Current Evaluation Practices

A general overview is presented of existing procedures used by Arkansas, Kentucky, and Massachusetts to monitor school improvement.

This section of the report presents a general overview of existing procedures used by Arkansas, Kentucky, and Massachusetts to monitor school improvement. These three states were chosen because Arkansas is modeling its Scholastic Audit procedures after the practices developed in Kentucky, including receiving training from their staff. Massachusetts was chosen because it has two different monitoring systems, as well as the highest National Assessment of Educational Progress (NAEP) scores in the nation (http://nces.ed.gov/nationsreportcard/states/profile.asp). These states appear to have the most comprehensive monitoring systems for school improvement in the nation, and together they seem to reflect the state-of-the-art in school auditing.

It should be clear from the onset that no state was located that had a statewide program evaluation of student performance.

It is important to note that no state in the nation has a statewide program evaluation of student performance - that is, a systematic program evaluation as described in Section 1 of this report. On the other hand, there are smaller, or local, program evaluations such as the Evaluation of Year One of the Achievement Challenge Pilot Project in the Little Rock Public School District (Barnett, Ritter,

Winters, & Greene, 2007).

Arkansas and Kentucky use a Scholastic Audit developed in Kentucky for monitoring school improvement. Massachusetts actually has two separate and distinct review processes.

Instead of program evaluations, Arkansas and Kentuckyuse a Scholastic Audit developed in Kentucky for monitoring school improvement (http://www.kde.state.ky.us/KDE/Administrative+Resources/School+Improvement/Scholastic+Audits+and+Reviews/), whereas Massachusetts actually has two separate and distinct review processes. As one part of its accountability system, the Massachusetts Department of Education oversees local compliance with education requirements through the Coordinated Program Review (CPR) (http://www.doe.mass.edu/pqa/review/cpr/). In addition, the Massachusetts Legislature created the Office of Educational Quality and Accountability (EQA) in July 2000, to provide independent and objective programmatic and financial audits of school districts in the state. The EQA is responsible to a five-member citizen council appointed by the governor. EQA's review process is called an examination, which is required of all schools identified as needing improvement and a random sample of successful schools. The CPR and EQA examinations are discussed after the presentation of the scholastic audits conducted in Arkansas and Kentucky.

**Scholastic Audits: Arkansas and Kentucky**

Arkansas initiated the Scholastic Audit, adapted from Kentucky, with 30 schools in the 2006-07 school year, and a similar number of schools are planned for the upcoming school year.

Arkansas initiated the Scholastic Audit, adapted from Kentucky, with 30 schools in the 2006-07 school year, and a similar number of schools are planned for the upcoming school year. The Scholastic Audit was authorized by the Kentucky General Assembly in 1998 (Ky. Rev. Stat. Ann. §158.6455). Section four of that statute directs the Kentucky Board of Education to establish guidelines for: 1) conducting scholastic audits, which include a process for appointing and training team members, reviewing a school's learning environment, efficiency, and academic performance of students; 2) evaluating each certified employee; and 3) reporting to the Kentucky Board of Education about the appropriateness of a school's classification (e.g., successful, needing improvement) and the assistance required to improve teaching and learning in the audited school (http://www.kde.state.ky.us/NR/rdonlyres/73738130-C2FE-4085-AFA8-9D7A5CE9AEB9/0/2513oversizepdf.pdf).

The Scholastic Audits conducted in Arkansas and Kentucky are *audits* or *reviews* rather than *evaluations*. Starting in the 2006-07 school year, schools in Arkansas designated as needing improvement for 3 or more years were required to participate in a scholastic audit conducted by the ADE.

As the name indicates, the Scholastic Audits conducted in Arkansas and Kentucky are *audits* or *reviews* rather than *evaluations*. The audits are intended to provide a very rigorous, in-depth assessment of factors that influence student learning in schools identified as needing improvement. Starting in the 2006-07 school year, schools in Arkansas designated as needing improvement for three or more years were required to participate in a scholastic audit conducted by the Arkansas Department of Education (ADE). According to the Arkansas Consolidated School Improvement Planning (ACSIP) handbook, these audits make recommendations to improve teaching and learning for inclusion in the comprehensive school improvement plan. According to the Scholastic Audit Guidebook (Fall, 2006), ADE audit site teams assess schools on nine standards and indicators for school improvement: 1) curriculum; 2) classroom evaluation/assessment; 3) instruction; 4) school culture; 5) student, family, and community support; 6) professional growth, development, and evaluation; 7) leadership; 8) organizational structure and resources; and 9) comprehensive and effective planning. Each standard, in turn, has 5 to 16 variance points (or dimensions) that are assessed by the site audit team using a four-category rating scale: 1) little or no development of implementation; 2) limited development and partial implementation; 3) fully functioning and operational level of development and implementation; and 4) exemplary level of development and implementation. Variance points are indicators where the ratings vary significantly between schools in need of assistance and successful schools. Examples of variance points from Kentucky are presented in Appendix A. The site teams that rate Arkansas schools on these variance points consist of six former teachers and administrators.

For comparison purposes, Kentucky conducts Scholastic Audits in a five percent random sample of successful schools to establish the "variance points" or factors that distinguish "successful schools" from those needing improvement. These factors serve as the criteria for rating schools on the nine standards and indicators for school improvement discussed in the third paragraph of this section of the report.

The Scholastic Audit Guidebook (Fall, 2006, p. 112) states, "Holistic scoring is the process of assigning a single performance level rating based on an overall view of an indicator. It is an inferential process in which the reviewer draws some overall conclusion based on specified criteria and standards about the

school's performance." The team arrives at a consensus regarding a rating of indicators of the "variance points" or dimensions of the nine standards listed above. For example, classroom instruction is indicated by several variance points, including, but not limited to varied strategies, alignment of strategies with goals, and alignment of strategies with learning styles.

This inferential process of holistic scoring yields qualitative data that are derived through professional judgments made from observation and reports from teachers, students, school officials, and parents.

This inferential process of holistic scoring yields qualitative data; that is, data that are derived through professional judgments made from observation and reports from teachers, students, school officials, and parents. Because ratings are based on professional judgments, the ADE hires experienced teachers and administrators as audit team members.

Evidence for ratings comes from several sources, including classroom observations; examination of displays of student work; and interviews with teachers, students, administrators, other staff, and parents.

Evidence for ratings comes from several sources, including classroom observations, examination of displays of student work, and interviews with teachers, students, administrators, other staff, and parents. Questions and rubrics are offered as guides to provide evidence for ratings. Each rating must be supported by three pieces of evidence. Other sources of evidence for these ratings include a stakeholder perception survey, school board policy review, and school leader self-assessment surveys. Prior to the site visit, each school must provide the team with a school portfolio, which consists of the ACSIP, curriculum alignment documents, district evaluation plan, district technology inventory, teacher lesson plans, master schedule, professional development activities, school handbook, school report card, school survey data, state assessment results, student achievement data, student work samples, samples of student assessments in core areas, local board of education policy manual, and writing portfolio analysis data.

These data, individually and collectively, provide an impressive array of evidence in support of the ratings assigned to the nine standards and indicators for school improvement listed in the Scholastic Audit Guidebook (Fall, 2006).

These data, individually and collectively, provide an impressive array of evidence in support of the ratings assigned to the nine standards and indicators for school improvement listed in the Scholastic Audit Guidebook (Fall, 2006), especially since the ADE requires three pieces of evidence for each rating. In tandem, these data provide a rigorous and thorough qualitative assessment of a school's practices, policies, and resources. In many respects a scholastic audit parallels or mirrors the process evaluation procedures discussed in Section 1.

This extensive scholastic audit is conducted over a period of three consecutive days. The purpose of the audit is to provide schools with an extensive array of actionable information relevant to school improvement efforts. It is an assessment - or thorough diagnostic inspection - of the deficits and resources that school officials need to

consider in devising and implementing their school improvement plans to achieve adequate yearly progress (AYP) as mandated by NCLB.

Within ten days of receiving a copy of the scholastic audit, school officials are to contact their school improvement supervisor to schedule a technical assistance visit. Each of these supervisors oversee about 15 to 20 school districts. Prior to this initial visit, school officials are to review the next steps and recommendations offered in the Scholastic Audit Report, and: 1) prioritize recommendations by year according to level of impact on teaching and learning; 2) initiate a review and analysis of the effectiveness of programs and services the school currently is implementing; 3) establish goals based on recommendations; and 4) identify high yield intervention strategies to improve student achievement and overall school performance.

According to the Implementation and Impact Guidelines of the Scholastic Audit, the school improvement supervisor should immediately schedule a visit to meet with school or district personnel (i. e., superintendent, principal, ACSIP chair) to discuss 1) the next steps, 2) the process for amending ACSIP, and 3) establishing a date for the next ACSIP meeting.

In follow-up ACSIP revision meetings, school officials and the school improvement supervisor: 1) incorporate the Scholastic Audit findings into the school's needs assessment; 2) expand the goal and benchmark statement; 3) review and identify existing ACSIP interventions and actions that would be appropriate to the recommendations made; 4) develop multiple sequential steps to implement and evaluate each new intervention; 5) identify roles and resources; 6) establish a timeline for implementation and evaluation; and 7) schedule a follow-up visit to provide technical assistance with identifying resources, professional development, and so on.

The ADE has developed a form (Implementation and Impact Check - Appendix B) to specify the steps to be taken in the ACSIP to implement the school improvement and to assess its impact. As seen in Appendix B, this form requires schools to report the status (not implemented, partially implemented, or implemented) of each action taken, whether or not that action has had an impact (yes or no), evidence for the impact, and nature of the impact. The types of evidence that can be presented are covered in the Scholastic Audit Guidebook, which may include teacher lesson plans, student work samples, student achievement data, and curriculum alignment documents.

ADE is in the process of developing an official schedule of when and how often follow-up ACSIP revision meetings occur, which will include an end-of-the-year follow-up meeting (beginning in the 2007-08 school year).

The primary emphasis in Arkansas and Kentucky has been on an extensive assessment or review of deficits and resources to inform school improvement plans to achieve adequate yearly progress.

In summary, the primary emphasis in Arkansas and Kentucky has been on an extensive assessment or review of deficits and resources to inform school improvement plans to achieve AYP. Beginning in the 2007-08 school year, the ADE is instituting follow-up procedures to monitor the progress of school improvement. These initial and follow-up audits or reviews consist of professional ratings based on observations, interviews, and reports provided by schools. With the exception of testing for student achievement, these data collected are qualitative. Efforts are made to establish the reliability and validity of these data by requiring a consensus among raters and three pieces of evidence for ratings assigned. However, it is important to keep in mind that these procedures, in final analysis, rely on professional judgments of particular auditors and school officials.

The Scholastic Audit procedures do not provide the systematic examination of statistical relationships between school improvement efforts (actions or programs) and outcomes or impacts such as student achievement.

The Scholastic Audit procedures do not provide the systematic examination of statistical relationships between school improvement efforts (actions or programs) and outcomes or impacts such as student achievement as discussed in Section 1. That is, while it is true that ADE monitors student performance in schools on criterion-referenced and norm-referenced testing, there are no statistical analyses of the relationships between educational interventions and student achievements (http://arkansased.org /testing/assessment.html). Conversations with ADE officials indicate that they are in the planning stages of formal program evaluations like those discussed in Section 1 of this report. ADE does disaggregate test scores according to the mandates of NCLB, and test scores are used to determine AYP and to identify schools and disaggregated groups that need improvement. Currently, an assumption is made that any gains in student achievement must be the result of any antecedent actions or programs that were implemented. However, as discussed in Section 1 multivariate statistical analyses would permit the ADE to begin to determine the effectiveness of actions and programs in changing outcomes such as student achievement.

A large volume of research indicates that student achievement gains are the result of many extraneous factors in addition to planned interventions, such as quality of teaching and characteristics of students, families, classrooms, schools, and communities.

As discussed in Section 1, a large volume of research indicates that student achievement gains are the result of many extraneous factors in addition to planned interventions, such as quality of teaching and characteristics of students, families, classrooms, schools, and communities (Odden, 2004; Odden, Borman, & Fermanich, 2004). Multivariate analyses (discussed in Section 1) identify the separate "effects" of each factor, including actions or programs, when they are considered simultaneously (or together in one equation). The reason NCLB requires disaggregation is because of the "effects" of extraneous factors on student achievement.

Analyzing interventions and other influences (e. g., parental income or motivation) together would also allow the ADE to determine the conditions and circumstances under which different interventions are effective. As discussed in Section 1, research literature indicates that interventions have to be modified under certain conditions (Odden & Wallace, 2006). Yet, no state has been identified that does multivariate analyses of relationships between interventions, extraneous factors, and outcomes. Without analyzing these relationships, there is no empirical evidence regarding the effects of specific programmatic interventions. This means states are not accumulating knowledge about what programs are effective with which students under what set of conditions.

### Examinations in the Massachusetts Office of Educational Quality and Accountability

There are two parallel reviews or audits in Massachusetts. The state department of education conducts a Coordinated Program Review (CPR), while an Office of Educational Quality and Accountability (EQA) carries out "examinations" under the auspices of a 5-member citizen Education Management Audit Council (EMAC) appointed by the governor.

Massachusetts, like Kentucky, also assesses both successful schools as well as schools that need improvement. As stated in the introduction to this Section 2, there are two parallel reviews or audits in Massachusetts. The state department of education conducts a Coordinated Program Review (CPR), while an Office of Educational Quality and Accountability (EQA) carries out "examinations" under the auspices of a five-member citizen Education Management Audit Council (EMAC) appointed by the governor. Historically, the department's CPR has focused on schools, whereas the EQA examination concentrates on school districts. The purpose of the EQA is to provide independent and objective programmatic and financial audits of school districts in the state. EQA calls its review process an "examination," and it assesses six accountability standards: 1) leadership, 2) curriculum and

instruction, 3) assessment and evaluation systems, 4) student academic support systems, 5) human resource management and professional development; and 6) financial systems and efficient asset management. According to the current EQA website, since 2002 EQA has reviewed over 150 districts, which is more than one-third of all the approximately 350 districts in the state. These districts include urban, suburban, rural, regional, and vocational-technical schools. To date EQA has reviewed all of the state's lowest performing districts, as well as all of the school districts in large cities (copies of the technical reports are located at: http://eqa.mass.edu/ reports/reports.asp.

All school districts receive a Tier I review annually, which consists of EQA examining disaggregated test scores in terms of levels of performance (e.g., proficiency) and consistency across time. Each year approximately 50-60 districts are then selected for further review and on-site visits by EQA staff. Those selected include: urban, suburban, and rural districts; regional, vocational, and single community K-12 districts. The majority of districts (60%) selected are "low" performing, or below the state average performance level on the student performance assessments. The remainder (40%) are selected at random; EQA is charged with reviewing all districts within the state.

The EQA website states that because the EQA team has no connection to a school district, candid interviews can be conducted at all levels within the district including: 1) the superintendent, 2) assistant/deputy superintendents, 3) business manager, 4) directors, 5) principals, 6) teachers, 7) district-wide program coordinators, 8) school committees, 9) municipal officials, and 10) the president of the local teachers' association. EQA team members, like the audit teams in Arkansas, must have extensive backgrounds in K-12 education. The difference is that EQA team members are responsible to an independent council instead of to the state department of education.

EQA team members, like the audit teams in Arkansas, must have extensive backgrounds in K-12 education. The difference is that EQA team members are responsible to an independent council instead of to the state department of education.

EQA prepares four different reports: 1) technical reports that provide in-depth ratings of schools on all levels of performance; 2) general reports which summarize findings of ratings for the school district; 3) annual reports that summarize statewide findings; and 4) research papers that examine longitudinal data on school districts. All technical reports that have been written for schools are available at this website: http://eqa.mass.edu/reports/technical.asp.

EQA makes every effort to schedule its onsite visit around the Coordinated Program Review conducted by the department of

education. A visiting team is typically composed of 5-7 examiners, who complete 5-10 reviews per year (http://eqa.mass.edu /resources/process3.asp). Prior to an onsite visit, EQA requests approximately 30 documents (e.g., school and district improvement plans, curriculum guides) for the period under examination (generally four years). EQA does not require any documents or reports that are not already required by the department of education (http://eqa.mass.edu/resources/docs/DocumentChecklist.pdf ). As a result of an extensive review of these advanced documents, a series of preliminary questions and concerns are generated in preparation for the onsite visit.

An EQA review is primarily a survey of management practices within the district rather than a scholastic audit.

An EQA review is primarily a survey of management practices within the district rather than a scholastic audit. Like the scholastic audits, however, the onsite visit is the centerpiece of the whole review process. This onsite visit typically lasts for four days. During this visit, EQA examiners meet with the majority of a district's administrators, including the superintendent, assistant/deputy superintendents, business manager, directors, principals, and district-wide program coordinators. Additional interviews are conducted with groups of teachers, the School Committee, the president of the local teachers' association, and municipal officials.

The standards and indicators used have been modified annually since the inception of EQA but have been consistently focused on the same six accountability standards (listed in first paragraph of this subsection). The most significant changes have occurred in the organization of the information and not to the standards or indicators being used as measures. Appendix C contains a copy of one standard (curriculum), findings, ratings, and evidence for each rating, of a school examined by EQA in 2005. The rating scale used by EQA is: 1) unsatisfactory, 2) poor, 3) satisfactory, and 4) excellent. Ratings represent a consensus of the EQA team, a procedure also used in Arkansas and Kentucky.

The rating scale used by EQA is: 1) unsatisfactory, 2) poor, 3) satisfactory, and 4) excellent. Ratings represent a consensus of the EQA team, a procedure also used in Arkansas and Kentucky.

After the EQA team leaves the district each examiner writes a report on the particular standard area they headed (an area in which they are specialized). This report details how the district performed relative to each of the indicators contained within the particular standard. Approximately one week later, the team reconvenes for a corporate session where each examiner presents their report back to the group. These draft reports are subjected to intense critique and scrutiny by the team. Edits are made based on the consensus opinion of the group.

The EMAC board meets regularly to review the reports generated from the EQA visits. The executive director of EQA presents the report to the board and highlights issues. The superintendent of a district reviewed has the opportunity to respond and raise questions or express concerns. The board then discusses the findings with the superintendent, after which the board votes on the report and makes a recommendation to accept or reject the report as is or with some action or modification.

The EMAC then transmits its findings and recommendations to the Governor, State Board of Education, Attorney General, President of the Senate, Speaker of the House of Representatives, and Clerk of the House of Representatives. The Clerk then forwards all materials to the Joint Committee on Education, Arts, and the Humanities. Copies of these technical reports are found at: http://eqa.mass.edu/reports/technical.asp. Each report is approximately 120 pages in length and contains an executive summary, overview of the EQA review process and the district, a Tier I analysis of student achievement and the Massachusetts Comprehensive Assessment System (MCAS) test data, the Tier II domain findings and summary, an explanation of proficiency index (PI), and the district's chapter 70 funding and net school spending history. Starting in school year 2005-2006, districts also receive an abbreviated (approximately 22 pages) general report, in addition to the technical report, that is oriented to the general citizenry. The content of these reports appears to be similar to the content in scholastic audits conducted in Arkansas and Kentucky, but with more emphasis on managerial aspects of school operations, instead of on scholastic features. The director of EQA reported that this agency is meeting to design a scholastic audit, with a particular focus on identifying instructional delivery mechanisms and strategies that distinguish between successful schools and schools that are placed on improvement lists (Rappa, 2007). The EQA also is in the process of designing program evaluations to determine the effectiveness of instructional programs. The director stated that EQA has realized that they need to move beyond reviews to evidence on what is working in terms of student achievement gains.

The content of these reports appears to be similar to the content in scholastic audits conducted in Arkansas and Kentucky, but with more emphasis on managerial aspects of school operations, instead of on scholastic features.

**Coordinated Program Review - Massachusetts Department of Education**

The Massachusetts Department of Education oversees local compliance with education requirements through the Coordinated Program Review (CPR). All reviews include monitoring for compliance with Title I fund use requirements. Significant aspects

In general, districts and charter schools that were in *identified for improvement* or *corrective action* status, for students in the aggregate or for student subgroups, and all districts with grant awards of $300,000 or more, received an onsite visit during the 2006-2007 school year as part of the monitoring process. Visits were made for the dual purposes of determining compliance and providing technical assistance, as needed, to improve program quality

of the 2006-07 school year Title I monitoring is accomplished through a desk audit of available data and documents. Districts and charter schools identified for improvement or corrective action are required to submit their written plans to improve student performance. In general, districts and charter schools that were in identified for improvement or corrective action status and all districts with grant awards of $300,000 or more, received an onsite visit during the 2006-2007 school year as part of the monitoring process. Visits were made for the dual purposes of determining compliance and providing technical assistance, as needed, to improve program quality (http://www.doe.mass.edu/pqa/review/cpr/).

**CPR Elements**

Depending upon the size of a school district and the number of programs to be reviewed, a team of two to eight department staff members, together with any necessary outside consultants, conducts a CPR over two to ten days in a school district or charter school. Each school district and charter school in the state is scheduled to receive a CPR every six years, with a mid-cycle special education follow-up visit three years after the CPR. Approximately 65 school districts and charter schools were reviewed in 2006-2007.

The CPR criteria for each program encompass the requirements that are most closely aligned with the goals of the Massachusetts Education Reform Act of 1993 to promote student achievement and high standards for all students.

Components of the CPR include:

- Review of documentation about the operation of the charter school or district's programs;
- Interviews with administrative, instructional, and support staff across all grade levels;
- Interviews with parent advisory council representatives and other parents;
- Review of student records for special education, English learner education, and career/vocational technical education. The department also selects a representative sample of student records for the onsite team to review, using standard department procedures, to determine whether procedural and programmatic requirements have been implemented;
- Surveys of parents of students with disabilities and parents of English learners. Parents of students with disabilities whose files are selected for the record review, as well as the parents

of an equal number of other students with disabilities, are sent a survey that solicits information regarding their experiences with the district's implementation of special education programs, related services, and procedural requirements; parents of English learners whose files are selected for the record review are sent a survey of their experiences with the district's implementation of the English learner education program and related procedural requirements; and

- Observation of classrooms and other facilities. The onsite team visits a sample of classrooms and other school facilities used in the delivery of programs and services to determine general levels of compliance with program requirements.

At the conclusion of the onsite visit, the CPR team holds an informal exit meeting to summarize its preliminary findings for the Superintendent or Charter School Leader and anyone selected by these school officials. Within approximately 45 business days of the onsite visit, the team leader forwards to the Superintendent or Charter School Leader a draft containing specific findings. The district then has 10 business days to review the report for factual accuracy. The final report is issued within approximately 60 business days of the conclusion of the onsite visit, and it is posted on the department's website at: http://www/doe.mass.edu/pqa/review/cpr/reports/.

The onsite team rates each compliance criterion on the following scale: 5- commendable, 4- implemented, 3- implementation in progress, 2- partially implemented, 1- not implemented, and 0- not applicable. Rating 3 is reserved for newly required programs, i.e., special consideration is given to implementation period. Where criteria are found "partially implemented" or "not implemented", the district or charter school must propose corrective action to bring those areas into compliance with the relevant statutes and regulations. This corrective action plan is due to the DOE within 30 business days after the issuance of the final report and is subject to the department's review and approval. The ratings are derived from "holistic scoring" in the same manner as described for scholastic audits. The review team discusses each criterion and arrives at a consensus concerning ratings. An example from approximately 200 criteria rated is presented in Appendix D from a final report found on the website (http://www.doe.mass.edu/pqa/review/cpr/reports/).

During phases of corrective action, the department staff will provide ongoing technical assistance. There are also limited special education technical assistance funds. School districts and charter

schools must demonstrate effective resolution of noncompliance identified by the department as soon as possible, but in no case later than one year from the issuance of the department's Final Program Review Report. Copies of these reports are found at the following website: http://www.doe.mass.edu/pqa/review/cpr/reports/.

The most impressive assistance offered to schools by departmental staff is the Performance Improvement Mapping (PIM). The PIM handbook and complete descriptions and instructions for the eleven-step self-review process are located on that website. PIM is a very thorough self-audit that is conducted by school staff through the facilitation of a Department staff member.

The most impressive assistance offered to schools by departmental staff is the Performance Improvement Mapping (PIM). The PIM handbook and complete descriptions and instructions for the eleven-step self-review process are located on that website. PIM is a very thorough self-audit that is conducted by school staff through the facilitation of a Department staff member.

The most impressive assistance offered to schools by departmental staff is the Performance Improvement Mapping (PIM) found at: http://www.doe.mass.edu/sdi/pim/. Located on that website are the PIM handbook and complete descriptions and instructions for the eleven-step self-review process. PIM is a very thorough self-audit that is conducted by school staff through the facilitation of a Department staff member. Schools are urged to appoint an audit leader other than the principal, who oversees the work of teams of teachers. Interdisciplinary teams are formed according to grades, and they meet regularly to proceed through the eleven-step process from data organization and assessment to data analyses and reporting. What seemed particularly impressive was the time and attention given to identifying problems and their "causes," and seeking solutions that have empirical support in the literature. All the various forms used to record information at each step of the process are also located on the website.

In conclusion, departmental officials related that the PIM process places considerable demand on school staff in terms of meetings and written documents. As a consequence, many schools have scaled back on their initial attempts to meet all the original expectations of PIM. However, the same officials stated that schools are finding the general process very helpful in identifying, prioritizing, implementing, and monitoring programmatic strategies.

## Conclusion

Taken together, the evidence assembled in the investigation phase of our report indicates that state departments of education monitor programs in schools that have been identified by student performance exams as needing improvement with rigorous audits and reviews. They are systematically gathering information on compliance with mandates from NCLB, and they provide disaggregated data on student performance. Our research found no state that systematically evaluates programs or examines relationships between programmatic interventions, other influences, and outcomes such as student performance. At best, the assumption seems to be that if student achievement gains are noted, they must be the result of antecedent programs.

Our research found no state that systematically evaluates programs or examines relationships between programmatic interventions, other influences, and outcomes such as student performance. At best, the assumption seems to be that if student achievement gains are noted, they must be the result of antecedent programs.

To accumulate knowledge about effectiveness of programs with different students under various conditions is to accumulate to inform policy decisions, schools and school districts must conduct program evaluations. However, Guthrie and Hill (2007, p. 7) observe, "School districts have not created data bases on which such decisions can accurately be made, and they are still constrained by politics and tenure laws from making purely performance-contingent decisions. When it comes to particular programs, districts are similarly ill equipped to measure effectiveness (not to mention cost-effectiveness) or to act on performance data. Though programs come and go and the stocks of schools in a particular locality change over time, transitions are caused more by funding availability and fashion than by judgments about effectiveness."

# References


Aarnoutse, C., van Leeuwe, J., & Verhoeven, L. (2005). Early literacy from a longitudinal perspective. *Educational Research & Evaluation, 11*, 253-275.

Barnett, J. H., Ritter, G. W., Winters, M. A., & Greene, J. P. (2007). *Evaluation of year one of the Achievement Challenge Pilot Project in Little Rock Public School District.* Fayetteville, AR: College of Education and Health Professions, University of Arkansas.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley.

Bradshaw, W., & Roseborough, D. (2005). Restorative justice dialog: The impact of mediation and conferencing on juvenile recidivism. *Federal Probation, 69,* 15-21.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Carmines. E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Thousand Oaks, CA: Sage Publications.

Childers, T. (1989). Evaluative research in the library and information field. *Library Trends, 38*, 250–267.

Collier, D. F. (2006). CDC grant program supports environmental health services delivery. *Journal of Environmental Health, 69,* 43-45.

Corcoran, K., & Fischer, J. (2000). *Measures of clinical practice: A sourcebook (3rd ed.; Vol. 2).* New York: Free Press.

Creswell, J. W. (2002). *Research design: Qualitative, quantitative, and mixed methods approaches.* Thousand Oaks, CA: Sage Publication.

Danielson, C. (1996a). *Enhancing professional practice: A framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (1996b). Teacher evaluation instrument. Retrieved July 31, 2006, from http://www.cesa11.k12.wi.us/Content/ProfessionalDevelopment/Initiatives/PI34/Pages/Danielson%20Rubric.pdf.

Danielson, C , & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

Devellis, R. F. (2003). *Scale development: Theory and application.* Thousand Oaks, CA: Sage Publications.

Freund, R. J., & Wilson, W. J. (1998). *Regression analysis: Statistical modeling of a response variable.* New York: Academic Press.

Gelman, A. (2005). Multilevel (hierarchical) modeling: What it can and can't do. Retrieved January 3, 2007, from, http://www.stat.columbia.edu/~gelman/research/unpublished/multi.pdf.

Greene, W. H. (1993). *Economic analysis.* New York: Macmillan.

Guba, E.G. (Ed.). (1990). *The paradigm dialog.* Newbury Park: Sage Publications.

Guilfoyle, C. (2006). NCLB: Is there life beyond testing? *Educational Leadership, 64,* 8-13.

Guthrie, J.W., & Hill, P.T. (2007). Making resource decisions amidst technical uncertainty. St. Louis, MO: Washington University, Daniel J. Evans School of Public Affairs.

Heckman, J. J., & Hotz, V. J. (1989). Choosing among alternative non-experimental methods for estimating the impact of social programs. *Journal of American Statistical Association 84,* 862-880.

Kalishman, S. (2006). Health program planning and evaluation: A practical, systematic approach for community health. *American Journal of Evaluation, 27,* 495-497.

Ky. Rev. Stat. Ann. § 158.6455 (2006).

Kimball, S., Milanowski, A. T., & Heneman, H. G. (2003, November). Research results and formative recommendations from the Study of the Washoe County Teacher Performance Evaluation System. Paper presented at the American Evaluation Association Seventeenth Annual Meeting, Sparks, Nevada.

Kimball, S. M., White, B., Milanowski, A., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79,* 54-78.

Mahoney, J. L., & Zigler, E. F. (2006). Translating science to policy under the No Child Left Behind Act of 2001: Lessons from the national evaluation of the 21st-Century Community Learning Centers. *Journal of Applied Developmental Psychology, 27,* 282–294.

Maruyama, G. M. (1998). *Basics of structural equation modeling.* Thousand Oaks, CA: Sage Publications.

Milanowski, A. T., & Kimball, S. M (2005). The relationship between teacher expertise and student achievement: A synthesis of three years of data. Paper presented at the American Education Research Association Meeting. Montreal, Quebec, Canada.

Milanowski, A. T., Kimball, S. M., & Odden, A. (2005). Teacher accountability measures and links to learning. In L. Stiefel, A. E. Schwartz, R. Rubenstein, & J. Sabel (Eds.). *Measuring*

*school performance and efficiency: Implications for practice and research (pp. 137-161).* Larchmont, NY: Eye on Education.

Milanowski, A. T, Kimball, S. M., & White, B. (2004). Teacher evaluation scores and student achievement: Replication and extension at three sites. Madison, WI: Wisconsin Consortium for Policy Research in Education, University of Wisconsin-Madison.

Netemeyer, R, G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications.* Thousand Oaks, CA: Sage Publications.

No Child Left Behind Act of 2001, Pub. L. 107-110, 20 U.S.C. § 6301, January 8, 2002.

Odden, A. (2004). Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education, 79,* 126-137.

Odden, A., Borman, G., & Fermanich, M. P. (2004). Assessing teacher, classroom, and school effects. *Peabody Journal of Education, 79,* 4-32.

Odden, A., & Wallace, M. (2006). *New directions in teacher pay.* Madison, WI: Wisconsin Center for Education Research, School of Education, University of Wisconsin-Madison.

Patton, M. Q. (2002). *Qualitative research & evaluation methods (3rd ed.).* Thousand Oaks, CA: Sage Publications.

Powell, R. R. (2006). Evaluation research: An overview. *Library Trends, 55,* 102–120.

Rappa, J. B. Personal communication. June 8, 2007.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models (2nd ed.).* Thousand Oaks, CA: Sage Publications.

Reichardt, S. W., & Borman, C. A. (1994). Using regression models to estimate program effects. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation (pp. 417-455).* San Francisco: Jossey-Bass.

Reynolds, A. J. (2005). Confirmatory program evaluation: Applications to early childhood interventions. *Teachers College Record, 107,* 2401-2425.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using sub-classification on the propensity score. *Journal of the American Statistical Association, 79,* 516-524.

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

Scholastic Audit Guidebook (Fall, 2006). Little Rock, AR: Arkansas Department of Education.

Shadish, W.R., Cook, T. D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis.* London: Sage Publications.

Streifer, P. A., & Schumann, J. A. (2005). Using data mining to identify actionable information: Breaking new ground in data-driven decision making. *Journal of Education for Student Placed at Risk, 10,* 281-293.

Weiss, C. H. (1998). *Evaluation: Methods for studying programs and policies (2nd ed.).* Upper Saddle River, NJ: Prentice Hall.

# Appendix A

**Kentucky Variance Points 2004-2005**

## STANDARDS AND INDICATORS FOR SCHOOL IMPROVEMENT

**Results vary greatly from Level 3's and Successful Schools**

**Standard 1 - Academic Performance - Curriculum**

Rigorous, intentional and aligned…

1.1a Aligned with academic expectation, core content, program of studies +
1.1b Discussions among schools regarding curriculum standards
1.1c Discussions among schools to eliminate overlaps, close gaps
1.1d Vertical communication w/focus on key transition points
1.1e Links to continuing education, life and career options +
1.1f Process to monitor, evaluate and review curriculum
1.1g Common academic core for all students *

**Standard 2 - Academic Performance - Classroom Evaluation/Assessment**

Multiple Evaluation and Assessment Strategies…

2.1a Classroom assessments are frequent, rigorous, aligned
2.1b Teachers collaborate in design of assessment, aligned
2.1c Students can articulate the expectations, know requirements
2.1d Test scores used to identify gaps +
2.1e Multiple assessments provide feedback on learning
2.1f Performance standards communicated and observable
2.1g CATS coordination - building and district +
2.1h Student work analyzed

**Standard 3 - Academic Performance - Instruction**

Instructional Program Engages All Students…

3.1a Varied instructional strategies used in all classrooms
3.1b Instructional strategies/activities aligned with goals +
3.1c Strategies monitored/aligned to address learning styles
3.1d Teachers demonstrate content knowledge *
3.1e Teachers incorporate technology in classrooms
3.1f Sufficient resources available #
3.1g Teacher collaboration to review student work
3.1h Homework is frequent, monitored and tied to instructional practice

**Standard 4 - Learning Environment - School Culture**

Effective Learning Community with Climate…

4.1a Leadership support for safe, orderly environment *
4.1b Leadership beliefs and practices for high achievement *
4.1c Teacher beliefs and practices for high achievement *
4.1d Teachers and non-teaching staff involved in decision making
4.1e Teachers accept their role in student success/failure
4.1f Effective assignment and use of staff strengths
4.1g Teachers communicate student progress with parents *
4.1h Teachers care about kids and inspire their best efforts *
4.1i Multiple communication strategies used to disseminate info +
4.1j Student achievement valued and publicly celebrated *
4.1k Equity and diversity valued and supported

**Standard 5 - Learning Environment - Student, Family and Community Support**

School Works with Families/Community to Remove Barriers…

5.1a Families and communities active partners +
5.1b All students have access to all curriculum
5.1c School provides organizational structure +
5.1d Student instructional assistance outside of classroom *
5.1e Accurate student record keeping system*

**Standard 6 - Learning Environment - Professional Growth, Development and Evaluation**

Researched-based, Professional Development and Performance Evaluation to Improve Teaching and Learning

**Professional Development**

6.1a Long term professional growth plans #
6.1b Building capacity with on-going PD *
6.1c Staff development aligned with student performance goals +
6.1d School improvement goals connected to student learning goals +
6.1e PD ongoing and job embedded +
6.1f PD aligned to analysis of test data #

**Professional Growth and Evaluation**

6.2a School has clearly defined evaluation process #
6.2b Leadership provides sufficient PD resources #
6.2c Evaluations and growth plans effectively used *
6.2d Evaluation process meets or exceeds statutes *
6.2e Instructional leadership needs addressed #
6.2f Leadership provides evaluation follow-up and support *

**Standard 7 - Efficiency - Leadership**

Instructional Decisions Focus On Support for Teaching/Learning, Organizational Direction, High Performance Expectations, Learning Culture, and Developing Leadership Capacity

7.1a Leadership developed shared vision
7.1b Leadership decisions are collaborative, data driven, performance +
7.1c Leadership personal PD plan focused on effective skills*
7.1d Leadership disaggregates data +
7.1e Leadership provides access to curriculum and data *
7.1f Leadership maximizes time effectiveness +
7.1g Leadership provides resources, monitors progress, removes barriers to learning *
7.1h Leadership ensures safe and effective learning #
7.1i Leadership ensures necessary SBDM policies
7.1j SBDM has intentional focus on student academic performance +
7.1k Leader has skills in academic performance, learning environment, efficiency +

**Standard 8 - Efficiency - Organizational Structure and Resources**

Organization Maximizes Time, Space, Resources…

**Organization of the School**

8.1a Maximizes organization and resources for achievement
8.1b Master schedule provides all students access #
8.1c Staffing based on student needs +
8.1d Staff 's efficient use of time to maximize learning *
8.1e Team vertical and horizontal planning focused on improvement plan
8.1f Schedule aligned with student learning needs

**Resource Allocation and Integration**

8.2a Resources used, equitable +
8.2b Discretionary funds allocated on data based needs
8.2c Funds aligned with CP goals #
8.2d State/Federal funds allocated with CP goals and data needs +

**Standard 9 - Efficiency - Comprehensive and Effective Planning**

School Improvement Plan…

**Defining the School's Vision, Mission, Beliefs**

9.1a Collaborative process

**Development of the Profile**

9.2a Planning process involves collecting, managing and analyzing data +
9.2b Uses data for school improvement planning#

**Defining Desired Results for Student Learning**

9.3a Plans reflect research /expectations for learning and are reviewed by team
9.3b Staff analysis student learning needs #
9.3c Desired learning results are defined

**Analyzing Instructional and Organizational Effectiveness**

9.4a Data used to determine strengths and limitations #
9.4b School goals are defined

**Development of the Improvement Plan**

9.5a School improvement action steps aligned with goals and objectives
9.5b Plan identifies resources, timelines & person responsible#
9.5c Process to effectively evaluate plan
9.5d Plan aligned with mission, beliefs, school profile, desired results

**Implementation and Documentation**

9.6a Plan implemented as developed
9.6b Evaluate degree of student learning set by plan
9.6c Evaluate student performance according to plan
9.6d Evidence to sustain the commitment to continuous improvement

* Highest Significance (Chi-Square .001) + Next Highest (Chi-Square .01) # And next Highest (Chi-Square .05)

**Appendix B**

ARKANSAS COMPREHENSIVE SCHOOL IMPROVEMENT PLAN (ACSIP)
AND FOLLOW-UP PROTOCOL

*IMPLEMENTATION AND IMPACT CHECK – Part I (Planning Guide)*
(TO EVALUATE, AMEND, AND UPDATE THE SCHOOL OR DISTRICT IMPROVEMENT PLAN)

**ArkansasEd.org**

**Overview:**

*Section 1117 of the federal No Child Left Behind (NCLB) Act of 2001 requires each state to establish a statewide system of intensive and sustained support and improvement for local education agencies and schools receiving Title I funds to increase the opportunity for all students to meet the state's academic content and achievement standards.*

**State Assistance to Schools –** NCLB requires that statewide systems of support include school support teams, distinguished teachers and principals, and provision of assistance from outside entities such as institutions of higher education, educational service agencies, or private providers of scientifically based technical assistance.

**State-Provided Technical Assistance for School Improvement –** In addition to the technical assistance that districts provide to schools in the improvement process, NCLB also envisions a role for states as technical assistance providers for both schools and districts that have been identified for improvement.

NCLB specifies particular types of assistance that the state should provide to schools and districts in need of improvement. Priority for state assistance is given to districts identified for improvement, followed by schools in corrective action, schools identified for improvement, and other Title I schools and districts.

**Services Provided by School Improvement Supervisors –** School improvement supervisors provide technical assistance to districts and schools in the development and implementation of the Arkansas Comprehensive School Improvement Planning (ACSIP) process:

- Data analysis
- Guidance for establishing annual measurable objectives
- Professional development
- Curricular and instructional practices
- Guidelines for program evaluation

- Locate and utilize community, state and federal resources for continuous school improvement
- Guidance on preparation and submission of plan amendments
- Participate in monitoring process

ARKANSAS COMPREHENSIVE SCHOOL IMPROVEMENT PLAN (ACSIP)

*IMPLEMENTATION AND IMPACT CHECK – Part I (Planning Guide)*
(TO EVALUATE, AMEND, AND UPDATE THE SCHOOL OR DISTRICT IMPROVEMENT PLAN)

## Scholastic Audit

**Arkansas Comprehensive Testing, Assessment and Accountability Program (ACTAAP) and the Academic Distress Program**

**9.12** Beginning with the 2006-2007 school year, schools designated in year three, four or five school improvement shall participate in a scholastic audit conducted by the Department of Education (or its designees).

**19.12.1** Results of the scholastic audit shall be presented to the superintendent within four (4) weeks of completing the scholastic audit. The audit shall make recommendations to improve teaching and learning for inclusion in the comprehensive school improvement plan.

Specific support provided to schools that have participated in an audit may include any of the afore mentioned technical assistance areas, as well as additional assistance for facilitating programs and interventions to correct audit findings:

- School findings and recommendations
- District findings and recommendations
- Expansion and implementation of school impact check
- Other (as applicable)

# ARKANSAS COMPREHENSIVE SCHOOL IMPROVEMENT PLAN (ACSIP)

*IMPLEMENTATION AND IMPACT CHECK – Part I (Planning Guide)*

(TO EVALUATE, AMEND, AND UPDATE THE SCHOOL OR DISTRICT IMPROVEMENT PLAN)

| District/School (Scholastic Audit) | ACSIP/School Improvement (Scholastic Audit) |
|---|---|
| Within ten (10) days of receiving their copy of the Scholastic Audit Report the school/district should contact their School Improvement Supervisor to Schedule a technical assistance visit.<br><br>**Prior to the meeting:**<br><br>✓ Review Next Steps and Recommendations.<br><br>✓ Identify and prioritize Recommendations (that will have the most impact on teaching and learning) by year – Year 1, Year 2 etc.<br><br>✓ Initiate a review and analysis of the effectiveness of programs and services the school/district currently are implementing.<br><br>✓ Establish goals as identified by each Recommendation.<br><br>✓ Identify high yield strategies needed to improve student achievement and overall school performance.<br><br>✓ Identify any additional resources needed for implementation of recommendations (Policies, Professional Development, Supplies/Materials, and Funds). | Upon being contacted, the School Improvement Supervisor should immediately (3-5 days) schedule a visit to meet with school/district personnel (Superintendent, Principal, ACSIP Chair) to discuss; Next Steps, Process for amending ACSIP, and Establishing a date for the next ACSIP meeting.<br><br>**Prior to the meeting:**<br>✓ Review school/district Audit report.<br>✓ Review school/district ACSIP<br><br>**Initial Meeting:**<br>✓ Access where the school/district are with implementation of Next Step.<br>✓ Ensure school/district has identified and set priorities in term of recommendations.<br>✓ Review process for amending ACSIP.<br>✓ Establish date and participants (Major stakeholders; including parents and America's Choice) for next meeting.<br><br>**Follow-up ACSIP Revision Meetings:**<br>✓ Add Scholastic Audit findings into Needs Assessment (Label Scholastic Audit Findings).<br>✓ Expand Goal and Benchmark Statement.<br>✓ Review existing ACSIP interventions and actions that would be appropriate to incorporate Scholastic Audit Recommendations.<br>✓ Identify appropriate interventions and actions to implement remaining Recommendations.<br>✓ Develop multiple sequential steps to implement and evaluate each new intervention.<br>✓ Identify roles and persons responsible, resources, etc.<br>✓ Establish timeline for implementation and evaluation. (Include evaluation action type.)<br>✓ Schedule follow-up visit to provide technical assistance with identifying resources, professional development, etc. |

*The ACSIP plan will be amended within thirty (30) days for immediate implementation of school/district's year 1 prioritized Recommendations. (Amendments requiring monies can not be made after March 31 of the fiscal year.) Implementation of recommendations requiring budget amendments must be made prior to the next school year.

*A follow-up status report of revised ACSIP should be submitted to the School Improvement Unit Leader within seven (7) days of revisions.

* The Implementation and Impact Check (Part II-page 6) should be completed by the school/district as applicable.

*Implementing Action Steps (Part 1 - B)*

| District: | School Name: | Supervisor: | Date: |
|---|---|---|---|
| Action Component: | | Standard: | |
| Priority Need: | | Goal: | |
| Causes and Contributing Factors: | | Objective: (Address the Priority) | |
| Current activities/programs that support this area: | | | |

# Action Plan

| District: | School Name: | Supervisor: | Date: |
|---|---|---|---|
| Action Component: | | Standard: | |
| Priority Need: | | Goal: | |
| Causes and Contributing Factors: | | Objective: (Address the Priority) | |
| Current activities/programs that support this area: | | | |

**Interventions:** (Initiatives or strategies to address the student academic, behavioral and social needs identified in the data analysis.)

| Action | Person Responsible | Timeline | Resources | Source of Funds |
|---|---|---|---|---|
| | | | | |

ARKANSAS COMPREHENSIVE SCHOOL IMPROVEMENT PLAN (ACSIP)

*IMPLEMENTATION AND IMPACT CHECK – Part II*

**Follow-up Protocol**

District:_____ Review Team:_____

School: _____ _____

Date:_____ _____

| **Action** | Status* | | | Has This Activity Had IMPACT (Yes) (No) | Evidence of Implementation | Report on Impact |
|---|---|---|---|---|---|---|
| Activity and Strategy | I | IP | NI | | List evidence | Impact on teaching and learning |
| | | | | | | |
| | | | | | | |
| | | | | | | |

*** I=Implement; IP=Implemented Partially; NI=Not Implemented**

## Appendix C
## Example of One Standard Rated by Massachusetts EQA

**Standard 5. CURRICULUM:** For the period of time under examination, the district, each of its schools, and programs utilized curricula that were aligned with the State Curriculum Frameworks in the core academic subjects of English Language Arts (ELA), mathematics, science and technology (and other tested core academic subjects as added). The curricula were current, academically sound, and clearly understood by all who administered and taught in the district.

**Preliminary Finding(s):**
- The district's curriculum was academically sound, and all who administered and taught it understood it.
- During the 2003-04 school year, staffing was inadequate to deliver the district's curriculum to all student populations.
- Modifications to the curriculum did not result in significant improvement in moving students from the 'Warning/Failing' and 'Needs Improvement' categories to the 'Proficient' and 'Advanced' categories on the MCAS test.

**Indicators**:
1. The district had written curricula for all grade levels and tested core content areas that were clearly aligned with the State Curriculum Frameworks.
**Rating**: Satisfactory

**Evidence**: The district's grades PreK-6 and grades 7-12 math and ELA curriculum guides had a consistent format and were aligned with the state curriculum frameworks. Interviewees said that the district had a five-year cyclical curriculum plan that reviewed and revised each curriculum and aligned it with the frameworks.

2. Each school in the district had a curriculum leader to oversee the use, alignment, quality, currency, and consistency of the district's curricula.

**Rating**: Satisfactory
**Evidence**: Administrators said that each school had curriculum leaders. Principals were responsible for curriculum and instruction at their schools. The district also had a Curriculum Planning Council, Curriculum Study Teams, and a grades K-12 science and technology coordinator. The junior high school and high schools had department heads for English, math, and history/social studies. Beginning with the 2003-2004 school year, there were grades PreK-2 and grades 3-6 coordinators in ELA and math and a social studies coordinator for grades K-6. Before that year, there were fewer coordinators who covered grades K-12. Coordinators had teaching responsibilities, and the district paid them a stipend for coordinator work. At staff meetings, grade-level meetings, cross-grade level meetings, department head meetings, and professional development activities, principals, department heads, and curriculum coordinators led discussions on aligning the curricula with the frameworks, test results, consistency in testing, and school-based issues.

3. The district had an established, documented process that involved teachers in the annual review and/or revision of curricula based on the analyses of results of standardized tests.

**Rating:** Satisfactory
**Evidence**: The district's teachers and administrators systematically reviewed the MCAS test data and updated curricula during the period under review. The district had a five-year cycle for curriculum review and revision. The document on which it was based, the grades PreK-12 Curriculum Review and Update Plan, was most recently revised in September 2003. The plan contained the district's mission, vision, curriculum goals, educational goals, and guidelines and the framework for curriculum review and update. The district had a curriculum planning council that identified curricular needs, developed a curriculum calendar, evaluated the curriculum, and monitored the work of the curriculum study committees. The council was composed of the superintendent, the assistant superintendent, the principals, guidance staff, department heads, and curriculum coordinators. The curriculum study committees for each discipline included department heads, elementary curriculum coordinators, grades K-12 teachers, and special education teachers. The committees' composition varied, but teachers were always included. The study committees met periodically with elementary and secondary staff to identify needs and priorities. Interviewees said that data were part of every curriculum review. The teachers looked at strengths and weaknesses, trends, patterns, and the areas of the curriculum in need of revision. They also looked at teaching assignments. This procedure corrected any redundancies or gaps in the curriculum. Interviewees said that they worked backward in their planning, asking themselves what students needed to know for college and high school. For example, expository writing was one area that colleges recommended for improvement. The high school added more expository writing opportunities for upper grade students. At-risk students in grades 7-12 needed more math instruction. The district added a course in applied mathematics in grades 7 and 8. The district added MCAS test preparatory activities for Grade 10 students and provided a tutor for students who failed the test in Grade 10. The district looked at the data, reviewed the committees' input, and changed the curriculum, programming, and support services as needed.

4. (In academic districts) The results of student assessment data (i.e., longitudinal, demographic, disaggregated, diagnostic, and/or surveys) indicated that the district implemented an established process to ensure the scope, sequence, and alignment of learning goals, competencies, and expectations from one grade to the next in grades K-12 in ELA, mathematics, science and technology (and other tested core academic subjects as added).

**Rating:** Satisfactory
**Evidence**: There was evidence that the district reviewed the results of the MCAS test and other tests. The reviews translated into a consistent and established process that ensured the scope, sequence, and alignment of learning goals, competencies, and expectations from grades PreK-12. These elements were an integral part of the district's curriculum guides. During the period under review, the district had a curriculum planning committee

that identified curricular needs, developed a curriculum calendar, evaluated the curriculum, and monitored the work of the curriculum study committees. The district also had a written plan for a five-year cycle of curriculum revision.
The district's math curriculum was aligned with the November 2000 math curriculum framework. The curriculum contained a scope and sequence chart that listed what should be taught and reinforced in each grade and the order in which the content should be taught. The content outline for each grade indicated the number of days to spend on each concept. For example, Kindergarten teachers were to spend eight days on sorting and classifying, and a Grade 6 teacher were to spend nine days on probability. The guides included learning standards and performance indicators, instructional methodologies, and strategies for assessment. The district's ELA curriculum was aligned with the state ELA curriculum framework. The curriculum contained the strands and standards that were to be taught, when they should be taught, and approximately how much time it should take to teach them. It also included skills and examples of instructional activities and assessment products for each standard.

5. The district's curricula in all tested content areas were aligned horizontally to ensure that all teachers of a common grade level addressed specific subject matter following the same time line, and vertically to ensure complete coverage, eliminate redundancies, and close any gaps.

**Rating**: Satisfactory
**Evidence:** The district's curriculum was horizontally and vertically articulated during the period under review. The district developed comprehensive curriculum guides, curriculum maps, and benchmarks to assure horizontal and vertical articulation. District staff and administrators met frequently throughout the period under review in grade-level and, cross-grade level meetings to eliminate redundancy, and in departmental meetings, curriculum study meetings, and vertical teams to ensure coverage. Teachers were aware of the curriculum content required for student mastery for purposes of continuity. The district used the MCAS test data to revise the curriculum to ensure coverage of subject matter. When testing showed that students did not score well on square root questions, there was a review of who should be teaching square roots and when. The analysis of test data and discussions of curriculum were ongoing and led to change.

6. Modifications to the curriculum resulted in improved, equitable achievement for all student populations.

**Rating:** Poor
**Evidence:** Changes to the curriculum did not result in improved achievement for all student populations. Abington's students with disabilities and its students eligible for free or reduced-cost lunch (FRL/Y) achieved Adequate Yearly Progress (AYP). Despite this achievement, the Cycle III Accountability Report indicated that the district's special education students did not meet their performance target. Less than one third of students with disabilities in Abington attained proficiency on the 2004 MCAS test. This percentage was significantly lower than that of regular education students in Abington. Abington's FRL/Y students performed slightly better than the state average in 2002 and

2003. In 2004, they scored at the state average. An analysis of overall results indicated little improvement in moving students from the 'Warning/Failing' and 'Needs Improvement' categories to the 'Proficient' and 'Advanced' on the MCAS tests in both math and ELE interviewees said that teachers identified at-risk students and supported them in the regular classroom.

Teachers referred needier students to the Teacher Assistance Team, which explored various approaches with the teacher. These included modifications to the curriculum, teaching methods, and materials. Title I support services were also available.
Students on IEPs had the services of special education teachers and paraprofessionals. Special education teachers co-taught in the classroom. Students followed a modified version of the district's curriculum. Special education students in grades 9-12 participated in Green Wave Cafe, an in-house work-study program.
The district gave staff training in Project Read, Telian Phonics, and differentiated instruction to meet the needs of diverse learners. At the junior high school, students had help in the regular classroom, a special education teacher with open periods during the school day, and ISSPs. The high school offered an MCAS test preparatory course. A tutor was assigned to students who had failed the MCAS test in Grade 10.

7. Staffing levels were adequate to deliver the district's curriculum to all students, as indicated by equitable rates of improvement for all student populations.

**Rating:** Satisfactory
**Evidence:** Interviewees indicated that staffing was adequate during the first two years of the period under review, but that the 2003-2004 budget cuts reduced staffing to inadequate levels. That year, the district cut approximately 23 full-time positions and seven part-time positions. During the 2003-2004 school year, class sizes in grades 3-12 increased. Interviewees said that the instructional program was not as effective with these higher student-to-teacher ratios. The high school lost an English teacher, a math teacher, a science teacher, and a social studies teacher. The school offered no electives and combined levels two and three of its three-level English course. The junior high school lost a social studies teacher and a foreign language teacher. One elementary school lost four classroom teachers and had as many as 33 students in a Grade 4 class and 34 in a Grade 6 class. The elementary schools also lost a Kindergarten literacy teacher, a part-time art teacher, and an elementary reading specialist. The district cut its curriculum leadership positions in music, art, and foreign languages. Interviewees said that the budget cuts affected Grade 4 so that fewer students scored in the 'Proficient' category on the MCAS tests. Interviewees said that the district restored many of the lost positions for the 20042005 school year.

8. The district established practices that adequately provisioned for and supported the curriculum and its overall effectiveness in all assessed subject areas and all levels.

**Rating**: Satisfactory
**Evidence**: Administrators said that the district provided what they needed to support a viable instructional program. Despite the 2003-2004 budget cuts, the district supported

professional development. The district continued to fund courses for the staff. Money was available for summer conferences, and stipends were available for summer curriculum workshops. The district used the John Collins writing program and trained all new teachers in it. An Alternative Learning Program was available to meet the needs of underachieving students and students with social skills issues.