Ganga Persaud, Department of Educational Leadership:
Clark Atlanta University.
Email: gpersaud@cau.edu
&
Trevor Turner, Department of Educational Leadership:
Clark Atlanta University.
Email: tturner@cau.edu

Evaluation designs in relation to No Child Left Behind Act. Examples from the field

Paper presentation at the

2008 National Evaluation Institute Conference,
October 9-11, 2008. The Hilton Wilmington Riverside: Wilmington, North Carolina

Abstract

The No Child Left Behind act (NCLB: 2001) requires that all students succeed at or above grade level, and appears to favor experimental designs in order to demonstrate effectiveness with respect to all students' learning. This would mean that effectiveness must be demonstrated by controlling for alternative explanations in the context of the learners. NCLB has forced the consideration of the contextual variables and necessitated the choice of an evaluation design based on the extent to which it could control for such variables. Hitherto, school systems and policy makers have ignored such variables. Though there is a preference for experimental designs, the relative merits of correlation and quasi-experimental designs are discussed when it is difficult to conduct randomization of subjects in the school settings. The methodology and results will be displayed to demonstrate how teachers conducted action evaluation utilizing time series as a quasi-experimental design. The method is particularly useful when parents might object to a correlation evaluation design.

Evaluation designs in relation to No Child Left Behind Act. Examples from the field

The evaluation issues

The evaluator's choice of a method for school evaluation would depend on the nature of the context, the planned program, the availability of models of evaluation, and the requirements of the No Child Left Behind Act (2001) that all students succeed at proficiency level.

There are three issues to be considered when conducting evaluation. First, how do we map the field to be evaluated? Academicians are dominated by the concern to define the various aspects of the evaluation field in terms of: (a) context, (b) inputs, (c) process, and (d) outcomes. These academicians publish in scholarly journals that are not read by practitioners such as school administrators and teachers (Tough, 2006). Second, how do we conduct evaluation so as to control for sources of errors in terms of selection when the student population is diverse? This aspect is rarely or inadequately demonstrated in evaluation textbooks, as it is more the domain of research. Third, should evaluation include the contextual variables at both the font-end and back-end of evaluation to determine the worth of planned input and processes as compared to the contextual variables in contributing to the outcomes? Several researchers indicate that the contextual variables explain more of the variance on students' standardized tests scores than school variables (Bernstein, Hess and shipman, 1961; Coleman report, 1965; Hart and Risley, 1995; and Rothstein, 2003). There is a need for evaluators to consider students' performance in relation to both school-planned variables and home contextual variables in order to ensure the design of school-planned variables to counteract the effects of home variables (Persaud and Turner, 2002, 2007). Based on their research, both Rothstein, 203, and Hart and Risley appear to suggest that it might require political policy to enhance home conditions, if the school is required to be successful in teaching low income students to learn to the same extent as middle income students. The No Child Left Behind Act (NCLB 2001) appears to favor outcome evaluation while ignoring the measurement of the input and process variables in terms of the relevance to the contextual variables. According to Tough (2006) it is an attempt to improve outcomes hoping that by doing so the social context variables will go away. Its strategy is to ignore the direct measurement of these variables for accountability purposes and to control for social context diversity through the use of scientific research. In the area of scientific research it has selected the experimental design. NCLB (2001) has identified the experimental method and randomized assignment of subjects as the preferred design to demonstrate causal relationships and to control for alternative explanations and possible sources of errors; though other techniques are recognized for correlated explanations (Neuwman, 2002). Therefore, as far as school evaluation for state accountability is concerned, outcome evaluation following an experimental design is the preferred approach as declared by NCLB (though this requirement is not being currently propagated or attained). These issues are discussed below.

Standard for evaluation

The standard of evaluation required by NCLB raises the issue about: (i) How to design evaluation for effective planning of teaching by administrators and teachers so that all students can learn; (ii) How to design the evaluation of the various programs to demonstrate validity in terms of all students learning, (iii) How to conduct evaluation studies to estimate the impacts of teaching variables in relation to home variables so that teachers could be guided to make curriculum choices so that all students can learn.. A brief examination of NCLB appears to suggest that administrators and teachers should design outcome evaluation based on experimental design for both regular teaching and implementation of special programs. However, it does not exclude quasi-experimental, correlation studies and even qualitative studies.

The No Child Left Behind Act (NCLB 2001) requires new testing and accountability for U.S. public schools, indicating that by 2014 all students will be proficient in mathematics and reading. The intention of the act is that African-American, Hispanic and the poor will reach 100 percent in both reading and math. NCLB is the reauthorization of the Elementary and Secondary Education Act, passed in 1965. It was observed that since 1965 the federal government spent over $300 billion to educate youth from low-income homes, though only 32 percent fourth graders could read at grade level. It would appear as from the following summary (U.S. Department of Education, 2005) that the NCLB intention is to increase accountability through:

1. Annual testing whereby states must conduct testing in grades 3-8. Further, a sample of fourth and eight graders is required to participate in the National Assessment of Educational Progress (NAEP) testing program every other year in the content areas of reading and math.
2. Academic progress whereby states must demonstrate that all students are performing at proficiency level by 2014. Each school is required to demonstrate Adequate Yearly Progress (AYP). If a school received Title I funding and failed to meet a AYP for two successive years, the state is required to provide technical assistance, and families must be allowed a choice of other public schools. A school that failed AYP for three successive years must provide supplementary services such as after-school tutoring.
3. Report cards for each school that indicate progress on state standards are required to be provided by each state
4. Teaching quality whereby the federal government will provide greater flexibility in spending federal funds to state districts that are earmarked for improving teacher quality
5. Reading first grants to provide for new competitive grants for states and school districts to conduct scientific and reliable research-based reading programs for all students to learn in Kindergarten through grade three. Priority for these grants will be given to school districts in high poverty areas.

Three possible evaluations are required:

(i) Annual yearly progress to demonstrate that all students performed at proficiency level: Report card on an annual basis
(ii) Teaching quality in terms of all students achieving at proficiency level as the basis for scientific evaluation of staff development programs that are provided under NCLB for improving teaching quality
(iii) Reading first grants where the school needs to demonstrate effectiveness with respect to scientific research, indicating the need for experimental, quasi-experimental, and correlation evaluation studies

The Department of Education (Neuman, 2002) defines a study as reliable research if it: (i) uses the scientific method and includes a research hypothesis, a treatment group and a control group, (ii) can be replicated and generalized, (iii) uses rigorous standards in its design, methods and interpretation of results, (iv) produces convergent findings using various approaches. These relationships are demonstrated in the following diagram (figure 1) There is an underlying implication that parents could choose schools based on the successive failure of an assigned school indicating a political pressure for schools to be productive or lose enrollment that could have consequences for school closure, and the rise of charter schools at the expense of public schools.

The act is preceded and followed by a plethora of federal, state and local reforms or programs. Many programs have emerged: curriculum standards, best practices programs, reforming curricula, expansion of preschool programs, charter schools, teacher evaluation, high-stakes testing, student assessment, school performance, high

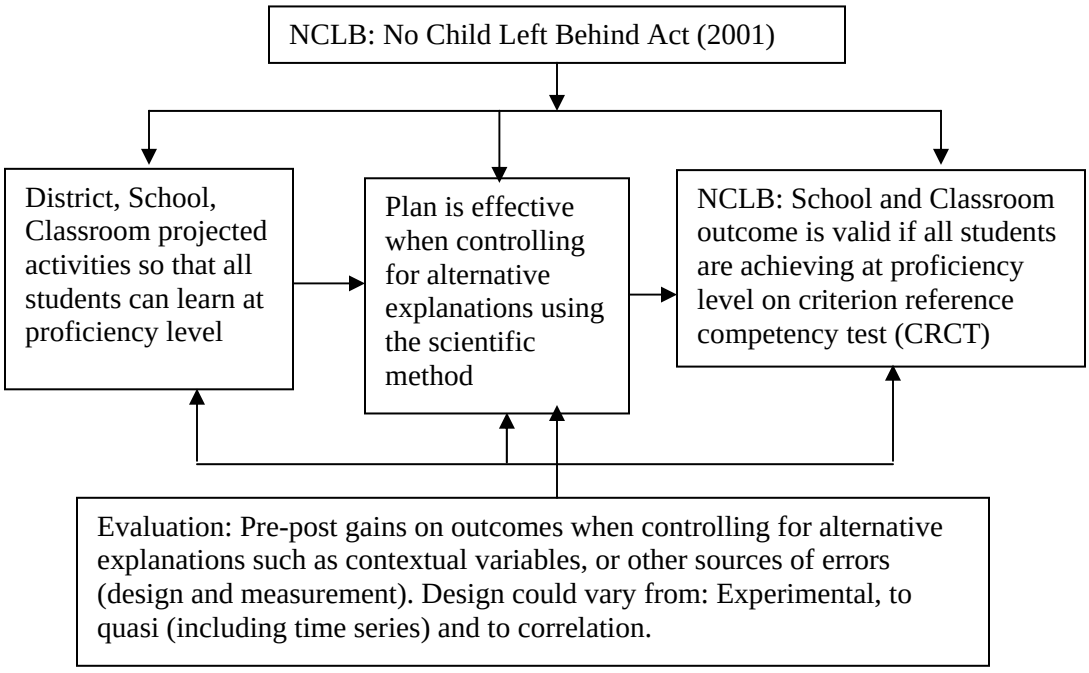

Figure 1: No Child Left Behind (NCLB) evaluation requirements

school graduation standards and revision of state and school funding formulas and others. The business sector has marketed several scripted programs (America Choice, Project Grad, Reading First, Success for All, etc.) to schools for enabling diverse teachers to teach in diverse schools and classrooms so that all students will learn the same state standards. These programs have not been effective as African-American, Hispanic and the poor as compared to White students who remained at least 20 points or two grade levels in both reading and math (Olsen, 2006). Tough, New York Times, 2006, reviewed the literature and raised two issues: What are the causes and what are the cures? According to Tough, academicians raise issues about causes in papers in scholarly journals. School principals who engage teachers and students in the teaching and learning process do not read scholarly journals and pay little attention to the research of academicians. In most instances, school reforms attempt to tackle the second issue (cures), independent from the first (causes). When private or charter schools are successful, they utilize radically different methods from the public school requiring greater commitment of resources than that obtained in NCLB. Further, there are two additional issues: The gaps between poor and 'better-off' children, and between Black children and White. Black children are three times as likely to grow up in poverty than White ones. Tough (2006) raised the issues as to whether the school alone could reduce the gaps in students' performance that are deep-rooted in socio-economic conditions, or whether political decisions should be made to remedy the socio-economic conditions as the basis for enhancing students' performance, or both. He argued that taken together it might be possible to enable the poor and minorities to achieve and bridge the gap, and overcome the deep-rooted early socialization of poverty. But, practitioners are unwilling to take on ethnic or socio-economic conditions, and concentrate on bridging the gaps on the academic side with the hope that if the academic gaps were eliminated, socio-economic and ethnic disparity would likewise disappear. Tough (2006) omitted to observe that practitioners are mandated to adopt these steps in the bureaucracy of the school as required by NCLB. Rothstein (203) and Hart and Risley (1995) provide data that indicate that even if the school and the community worked together it might not be possible to eradicate academic differences. The economic variables are too strong and both political and economic structures are required to support the education of poor parents' children. The fall of communism appears to suggest that it might not even be possible to eliminate socio-economic differences. Economic differences are based on the industrial stratification that requires different but hierarchical knowledge, skills and dispositions to function for mass production purposes, and the school by definition must be in correspondence. However, based on the philosophies of Dewey (1916) and Freire (1970 & 1973), if the political system and schools wanted all students to achieve at proficiency level, then it might be possible if students were taught at their individual baseline and on a growth evaluation model of testing in the context of their social and occupational experiences. In such an event, a plan and planning would need to identify the student-baseline performance and the causes for failure. The program of activities would be designed to cope with the causes, and evaluation would be designed to determine if the plan eliminated the contributions of the causal variables. Next, we turn to a systematic review of evaluation approaches to estimate their relative utility in conducting evaluation relative to NCLB.

Categorizing evaluation models and relevance to NCLB

There are several categories of evaluation approaches or models (Worthen and Sanders,1973; Gredler, 1996; Kilpatrick, Sanders and Worthen, 2004; Chen, 2005). These authors have not categorized evaluation models within a systems framework. However, the theoretical frameworks outlined by Tyler (1949), Stufflebeam (1973), Cunningham (9182) and NCATE 2000 and systems management (Mescon, Albert and Kedouri, 1988) appear to map the planning field in terms of a context-input-process-output framework. Evaluation models could be categorized by the extent to which they refer to the product or outcomes, process, inputs, or context, or all of the components and collaboration. In addition, there is a need to insert the need to research the context as part of both planning and evaluation We have articulated the several models of evaluation within a research and systems framework as shown in the following diagram (Figure 2). Within this framework:

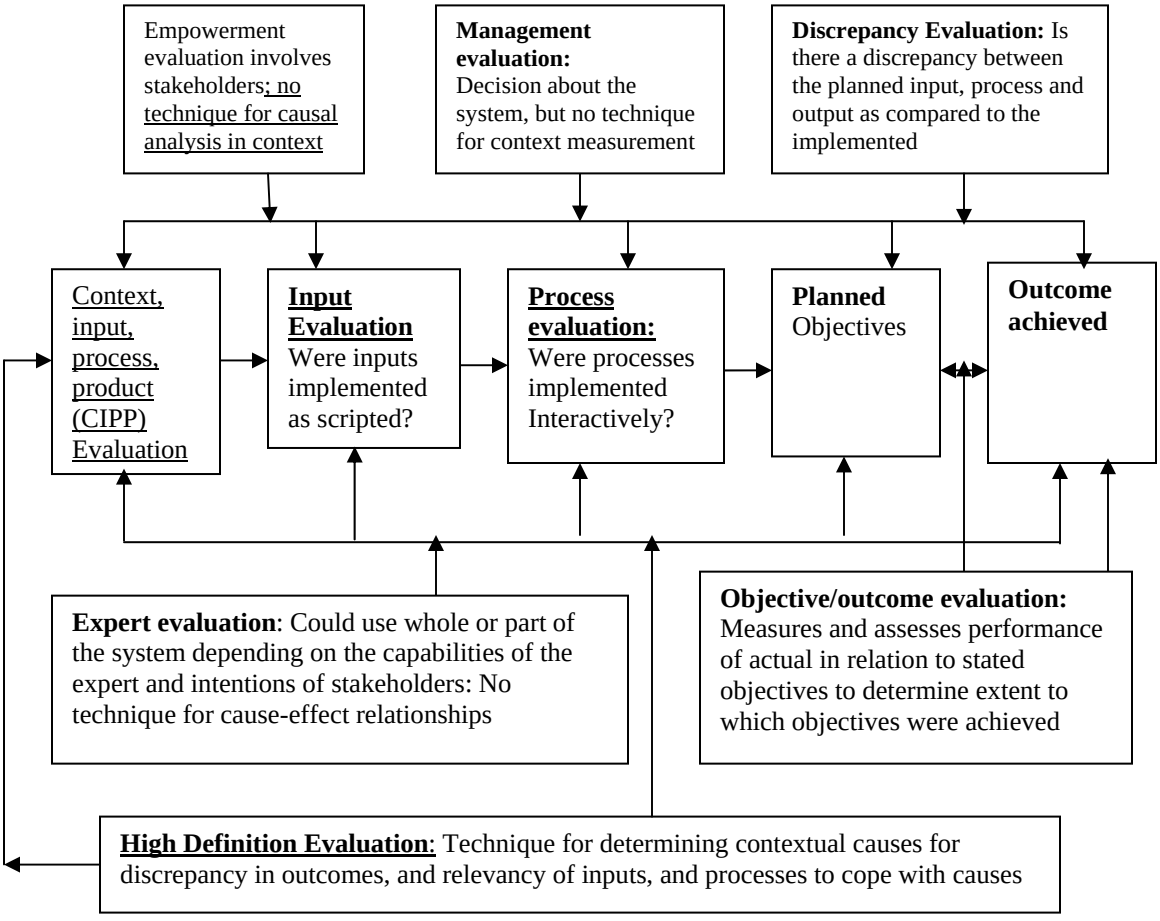

Figure 2: Evaluation orientations in relation to a comprehensive mapping of the evaluation field

Objective or outcome oriented evaluation is the process of collecting and analyzing data to determine the extent to which the objectives as defined are achieved in terms of actual outcomes. This evaluation process utilizes objectives or outcomes as the criteria for determining effectiveness of program activities. McGregor (1960) and Drucker (1964) synthesize the following essential steps: formulation of objectives, activities to implement the objectives, systematic evaluation of performance outcomes to determine effectiveness and taking corrective action based on evaluation results. Metfessel and Michael (1967) utilize the objective aspects of Tyler's curriculum model to focus on the outcome side of evaluation. If the objectives are not met in terms of measured outcomes, the program or teacher plan is deemed ineffective and discounted. Odiorne (1965) developed MBO whereby managers and subordinates set goals and use the goals as a basis for evaluating the effectiveness of the input activities. (Chen, 2005, pp. 202-203) suggests that the appropriate design to control for alternative explanations is the experimental randomized design. This is in support of the NCLB experimental design. Based on this model, a school would need to state pre-treatment results on outcomes as indicated on state proficiency tests. Next, the school would need to randomly assign students to the treatment and control group so as to control for contextual variables such as gender and SES variables. A Post-test is conducted to determine the effectiveness of the treatment.

In practice, it is difficult to conduct random assignments of subjects in a school as recognized by the State Department (Neuman, 2002). Schools produce group data for two or three successive years, and there is no way to control for the effects of gender and SES. Further, there is no way to examine whether the standardized curriculum and standards were in alignment with the diversity of students and might have contributed to variance in students' performance. In addition, NCLB allows each state to set its own testing standards so that each state could set a proficiency level it could achieve (Rorthstein, 2003). NCLB requirements of experimental designs to control for alternative explanations in terms of contextual variables have not been incorporated. This aspect is not enforced by NCLB.

Process oriented evaluation is the process of collecting and analyzing data to determine the extent to which action steps as defined on some instrument are implemented in the interaction process. This evaluation process is seen in teacher evaluation instruments that measure interaction in the observation of the teaching process. Flanders interaction analysis (1970), and state evaluation instruments are examples of such instruments. Teacher behaviors are specified on such instruments, and once the teacher performs them as observed, the teacher is deemed as meeting expectation or target. There is no attempt to consider the teacher behaviors in relation to the diversity of students in terms of gender or ability level, and the need to make adjustments in terms of diverse social experiences. No consideration is given to the outcomes in terms of students' learning on the full range of the Bloom's taxonomy in relation to the social context of the learners. NCLB requirements of experimental designs to control for alternative explanations in terms of contextual variables have not been incorporated.

Input program evaluation is the process of collecting and analyzing data to determine the extent to which the program action steps (to achieve the outcomes) are implemented as defined. The degree of implementation is considered as evidence of effectiveness. In

the best practices programs, administrators are involved in simply monitoring the program in the implementation phase to determine if the program was implemented as designed in the scripted program manual. The behaviors of the teachers are effective if they conform to the scripted program. No consideration is given to the outcomes in terms of students' learning in relation to the diverse social context of the learners, standardized tests, or the capability of the teacher. As a result many reading first programs were considered effective because the schools implemented them as planned. However, there was not the same impact on students' outcomes (Gold, 2008). NCLB requirements of experimental designs to control for alternative explanations in terms of contextual variables have not been incorporated. It should be observed that in the real world for the use of medication, the FDA requires a manufacturer of a drug to demonstrate experimentally the effectiveness of the drugs in terms of curing the patient as prior requirement of approval for sale to consumers. In the same way, reading first developers and others ought to produce experimental results for prior approval to the state department of education before introduction to schools.

Discrepancy oriented evaluation is the process of collecting and analyzing data to determine the extent to which the program activities, process and objectives as designed were implemented in practice (Stake, 1967; Provus, 973; House, 1980). In Stake's countenance model the evaluators map the field of the program development and intentions in terms of: rationale, inputs, transactions and outcome and determines there is congruence between each area of the intended and the actual practice. The contextual variables are ignored. NCLB requirements of experimental designs to control for alternative explanations in terms of contextual variables have not been incorporated.

Management or systems evaluation is the process of collecting and analyzing data to determine the extent to which the inputs are related to the process and both are related to outcomes (Mescon, Albert and Kedori, 1988). This is a requirement of NCATE (National Council for the Accreditation of Teacher Education, 2000). NCATE stipulates that assessment as standard II be used to measure Outcomes as Standard I in order to utilize such data to plan for clinical experience (Standard III), Diversity of the curriculum (Standard IV), Faculty development (Standard V) and Leadership and resources (Standard VI). In this system, assessment and evaluation are used in a cyclical process overtime to facilitate a "best fit" alignment between input and outcome variables. This is the domain of research. It is a four step model: Assessment to set baseline to plan, planning of standards III to VI, evaluation to determine effectiveness of outcomes in relation to plan, and to offer feedback for change, and the use of a continuous cycle over time (at least three years) for adjustment in contextual variables. Assessment and evaluation are focused on outcomes and not the contextual variables that might contribute to variance in outcomes. The assumption is that contextual variables would be controlled by time and not an experimental design. In a sense, it is an outcome evaluation following the requirement of three-year comparisons of outcome data. But it becomes a management system because of its cyclical use for making decisions about change in the input variables.

Context-input-process-product (CIPP) evaluation (Stufflebeam, 1973) is the process of collecting and analyzing data to determine the extent to which the context, inputs, process and product are inter-related in an alignment system. It differs from the management system because it considers the context as prior to and as influencing the inputs, process and product. It refers to causal effects, but has no technical method for measurement that would inform the management of the system. In the best practices programs, administrators would need to examine the extent to which student outcomes could be explained by the teacher delivery process in terms of classroom interactions, as observed, the planned inputs, and the social context of the students. That is to say, do the input and delivery process in the same classroom explain the students' outcomes as compared to the contextual factors? Tyler (1949) provides a precise method of evaluating objectives as outcomes, but also argues that objectives should meet the social, psychological and environmental needs of learners. It is a useful method in qualitative methodology (Patton, 1997). However, there is no precise method of identifying and measuring the contextual factors as predictors of students' performance.

Empowerment evaluation is the process of collecting and analyzing data to determine the extent to which the evaluator facilitates the personnel and stakeholders in decision-making about the evaluation steps as the basis for enhancing growth and commitment of all concerned as well as effective evaluation **(Fetterman, 2000)**. However, the method of defining the planning and program field and conducting the evaluation in correspondence does not appear to follow the systems process of context-input-process-outcome evaluation. It is possible to have participation while not making efficient decisions, because there is no method for guiding participants in making effective choices in terms of a system approach. A school that desires to utilize this process could get bogged down in the collaboration process that could displace the outcome evaluation required by NCLB

Expert evaluation is the process of utilizing experts to map the evaluation field and to utilize their expert knowledge about the required standard to determine the extent to which the current practice approximates the standard. This is the approach of the various accrediting organizations for schools. This process of school evaluation for accreditation has not led to more effective schools. The main reason is the non-alignment of each student performance to his/her gender and socio-economic conditions as well as to the teacher methodology and curriculum standards.

High definition evaluation system (Persaud and Turner, 2002, 2007) is the process of collecting and analyzing data to determine the extent to which a program plan is defined to counteract the causes of a problem as the basis for determining effectiveness with respect to outcomes. In terms of NCLB, if there is a variance in students performance, then evaluation should be conducted to determine the extent to which the variance could be explained by the input and process variables and/or alternative explanations such as: the social context variables of the students, etc. For example, in the best practices programs, administrators would need to examine the extent to which the student outcomes could be explained by the teacher delivery process in terms of classroom interactions as observed, and the planned inputs in relation to the social context variables

of the students. That is to say, do the input and process variables demonstrate that they counteracted the causes of the learning problems in terms of students' different social and learning backgrounds as the basis for improving students' outcomes. This approach appears to support the NCLB research requirement.

According Cartwright (1973) if there were no problems, there would be no need for planning, and for a problem to be solved the solution must match the definition of the problem.

If P (Problem) is defined as P = P (X1, X2, X3, .. Xn)

and if P = S, then, Solution S = S (X1, X2, X3, .. Sn)

However, the model does not provide a path for obtaining the best possible definition of the problem, and, hence, the solution could be wrongly selected. This is also a problem with strategic, and other models of planning. How are we to know that P is accurately defined? Dewey (1939) identifies a means-end relationship in any plan or problem situation and suggests that the means-end relationship is identical to the part-whole relation where the whole is desired and the parts are causally related. Further, his philosophy states that to desire a cause independent of an effect does not make it a means; to desire an effect independent of the means does not make it an end; rather it is the desirability of both in their entirety and in relation to one another that makes one a means to the other. Dewey's philosophy indicates that the key to defining a problem is to identify the means-end relationship in a given situation. This is the identical approach to the definition of a problem in scientific research where independent (causal) and dependent (outcome) variables are used to define a problem (Kerlinger, 1992). Following Dewey, Ishikawa (1976) used a fishbone technique to analyze a problem in terms of its sub-parts (as in spine and sub-spines of a fish applicable in quality circles). Deming (1986), following Ishikawa, uses a cause-effect analysis similar to Dewey. This contrasts with strategic planning, etc. That is to say, in order to examine why an end exists one must examine the causes. The causes explain the end. Persaud and Turner (2002) synthesized the literature to indicate that:

If a problem P(failure) = P(C1, C2, C3, C4...C5), where C1 to Cn are the causes.

Then, since P = S when the solution is successfully solved (Cartwright, 1973),

S = S(C, C2, C3, C4... Cn)

The model predicts that the desired objective will be achieved if the solution counteracts the causes of the problem. If a problem was not solved, the model tells us that:

1. The problem was not appropriately defined in terms of the causes,
2. The solution was not selected to counteract the causes of the problem, or both.

In planning, the solution must be selected to counteract the causes in order to solve the problem. Therefore, evaluation ought to consider the extent to which: (a) The outcomes were achieved, (b) The inputs and process variables counteracted the causes for low

outcome performance as observed in the pre-test condition, (c) The inputs and process variables contributed to gain in outcome performance, and that the effects of the causal variables in the pre-planning context were minimized. In terms of NCLB, students' baseline performance must be disaggregated by each student gender, and SES variables (principally home verbal style, academic and economic support systems) and by each teacher as the basis for determining the choice of curriculum and teaching, and evaluation strategies. In this way, evaluation can serve to determine the extent to which the effects of home factors have been counteracted by the choice of curriculum, teaching and evaluation strategies.

Operational steps for High Definition Evaluation: Correlation design

Based on these considerations, a High definition Evaluation (HDE) system is defined as the extent to which the program as a system has a research method for determining the causes for failure to achieve outcomes as the basis for choosing and aligning program activities to counteract the causal variables in a pre-implementation condition (Pre-Test data), and whether formative evaluation is utilized to ensure continuous process alignment so as to ensure that in a summative evaluation, the program activities improved outcome performance while the effects of causal variables had little or no effects. Cronbach et al (1980) suggest that contextual variables could influence the results. Therefore, it is necessary to determine how the contextual variables function to minimize the effects of program strategies. Patton suggests qualitative strategies to provide in-depth understanding about process variables. Based on these considerations, each teacher, in order to enable all students to learn as required by NCLB, would need to be supervised to function as follows:

1. Identify the range of students' performance on the state's standardized test (Pre-test or pre-implementation of treatment outcome data)
2. Determine each student's gender, family type (single or two parents, etc.) and identify possible causal relationships with variances in 1
3. Construct lesson plans in terms of teaching for higher order thinking skills in relation to students' contextual experiences so as to counteract the causal variables in 2
4. Teach the lessons
5. Construct tests to assess performance on a weekly or bi-weekly basis and to conduct change as required
6. Utilize post-treatment data on students' performance on the state's standardized test to compare with pre-test to estimate gains on each student in terms of NCLB
7. Conduct summative evaluation to estimate growth in students' outcomes in a pre-post comparison in relation to social context variables in 2.

In a case study, an eighth-grade teacher was asked to practice the above steps. The teacher was asked to rank order the students' performance pre-treatment ITBS score in eighth grade language arts, and to record the corresponding gender, single and two parents family type, occupation, ad race of each student. The teacher noted that the occupation of the parents more than the other variables was related to the pretest student performance. The treatment was conducted. The teacher utilized explanations,

questions and feedback processes in relation to students' experiences for teaching for higher order thinking skills on the full range of the Bloom's taxonomy. Tests were constructed to measure the growth on higher order thinking skills on a bi-weekly basis so that students' and the teacher utilized the data for making improvements. In a regression analysis of the data the posttest Language Arts scores on the ITBS were not explained by gender, family type or race. Instead, the occupation of parents significantly but inversely predicted the variance in the scores. Inspection of the raw scores in relation to occupation clearly supported the statistical analysis that students of the low occupation parents made greater gains than students of the higher occupations (Persaud, 1996). It would appear that each teacher could counteract the social contributors to students' academic performance, if teaching strategies are conducted to teach higher order thinking skills in correspondence with students' experiences, and the teacher assessed students' performance and made adjustments in alignment with standardized tests.

The authors trained 8 elementary school administrators in the above steps. They were asked to supervise grade chairs to work collaboratively with their respective teachers in conducting the steps. When the administrators were observed in the field, they were unable to practice the behaviors because teachers lacked knowledge, skills and dispositions to: (a) research the causal variables of student performance, (b) plan lessons to counteract the causal variables, (c) teach higher order thinking skills in relation to student social experiences, (d) evaluate the teaching process in terms of the technical process required to utilize students' experiences in the teaching higher order thinking skills, (e) construct tests for assessing higher order thinking skills on the content they have taught. We have inferred from our experiences that these are the dominant blockers in terms of teaching so that all students can learn.

Schools and teachers tend to depend on benchmark tests that are administered very six weeks while the results tend to have a four-week wait time. Feedback from such tests is rarely utilized to inform teaching. The administrators are now involved in training their grade chairs as trainers of teachers. One administrator who practiced the process was promoted from the elementary to the middle school as principal. The parents whose children were moving from the elementary to the middle obtained her transfer.

We turn to the use of various forms of experimental designs for controlling for contextual variables that rely on measuring outcomes while attempting to keep the contextual variables constant

Possible quasi and experimental designs

Time series designs are quasi-experimental designs that would control for alternative explanations of the treatment in a form that would be valid in terms of the requirement of NCLB. The unstated intention of NCLB is to control for alternative explanations without identifying them. In time series designs, there might be one or two groups with several

pre-test scores (01 to 04) before the treatment that teachers can easily implement daily without supervision.

In a single group case, a teacher in a classroom could teach traditionally for four weeks with a test at the end of each week (O1 to O5) as the pretest. The treatment is then conducted followed by several post-test scores (05 to 08) as shown in the following diagram (figure 3). The teacher could observe trends in the scores with respect to the traditional and treatment methods. If the tends in scores indicate an upward rise during the treatment more than that of the traditional, the treatment could claim the increasing scores. The mean scores of the treatment (O5 to O8) could also be compared with the traditional (O1 to O4). In this case, the class becomes two groups with pre-post tests respectively. The two groups are the same in every respect except for the treatment therefore gain scores could be attributed to the treatment. There is still, however, the question that natural maturity could explain the gain.

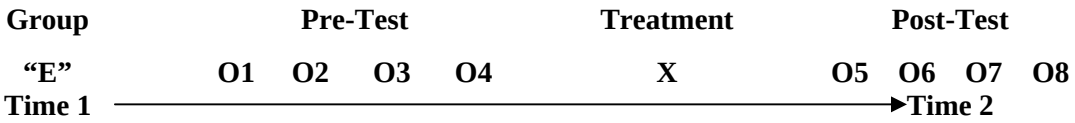

Figure 3: Single group time series

We gave administrators in a doctoral class on evaluation to conduct with a teacher the exercise of having the teacher to teach traditionally and conduct successive tests at least three times, then to teach innovatively and test at least three times. The first set of scores could be considered a pre-test mean score for the traditional method and the last three tests as the mean posttest score for the innovative. Gain scores would reflect effectiveness for the innovative method. The exercise became impractical because the principals and assistant principals and teachers could not construct MCQ tests based on content they were teaching. Further, benchmark tests were conducted every six weeks and the results took three or four weeks to get to the school. The standardized CRCT was conducted annually. The principals and assistants were taught to develop MCQ on content taught. They were asked to conduct a study (writing in progress) in which they worked with a grade chair to conjointly plan a lesson, teach the lesson (video-taped) and construct MCQ to evaluate the lesson. They were expected to supervise the grade chair in continuing the process and train fellow teachers in the process. In an end project interview with the authors only two were able to utilize the process with success.

The one group might not satisfy the issue of natural maturity to meet rigorous scientific requirement. To control for sources of errors such as history, maturation, selection, etc., we need to use a second group with no treatment as shown in following (figure 4). This would require coordination between teachers that would distract from preparation, and introduce inter-personal conflict variables. The method is operationally difficult in a natural school environment.

Group	Pre-Test				Treatment	Post-Test				Post-Pre-Test
"E"	O1	O2	O3	O4	X	O5	O6	O7	O8	(Gain)
"C"	O1	O2	O3	O4	NO	O5	O6	O7	O8	(Gain)
Time 1 →										Time 2

Figure 4: Two groups time series

Quasi-Experimental design as an alternative

When the administrators find it difficult to orchestrate random sampling and assignment of subjects in a true experimental design, they may want to attempt the quasi-experimental design. In this design, the subjects are reasonably matched by grade level. Students in classes at the same grade level are supposed to be fairly homogeneous in terms of social characteristics. Therefore, it might be possible to select a reasonably matched control group, and administer a pre-test on the outcomes of both groups. The treatment is conducted on the experimental only, and a post-test is administered on both groups as shown in the following diagram (figure 5). The two gain scores are compared. If the experimental shows a significant gain, it can be assumed that the gain was due to the treatment and not to the contextual variables. This claim can be made on the basis that though the pre-test baseline might vary, the comparison of the gain score minimizes the effects of such differences.

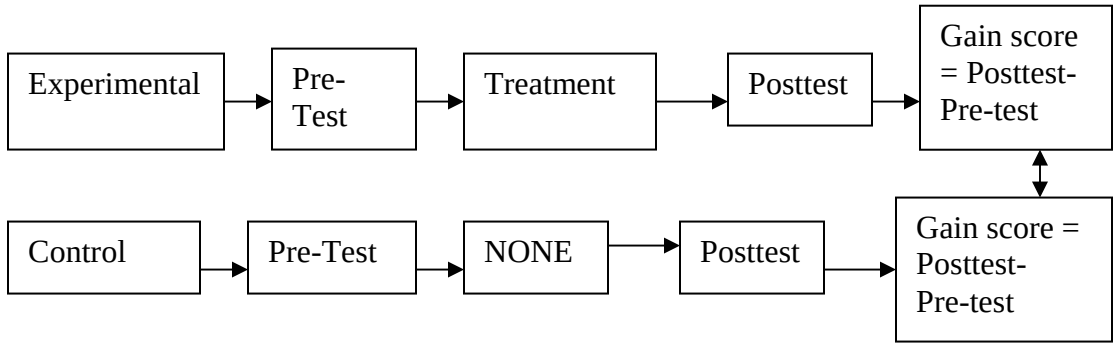

Figure 5: Quasi-experimental design using non-random selection and pre-post test

True-experimental design

For the scientific researchers, the quasi-experimental design does not fully control for contextual variables. They prefer, the true-experimental design in which the random assignment of subjects is critical for controlling for sources of errors. These relationships are demonstrated in the following diagram (figure 6).
The experimental design fits the intention of NCLB neatly, but it is difficult to implement on a daily basis in a school environment. It requires up-front planning. In school organizations practitioners hardly ever consider the random assignment of students into

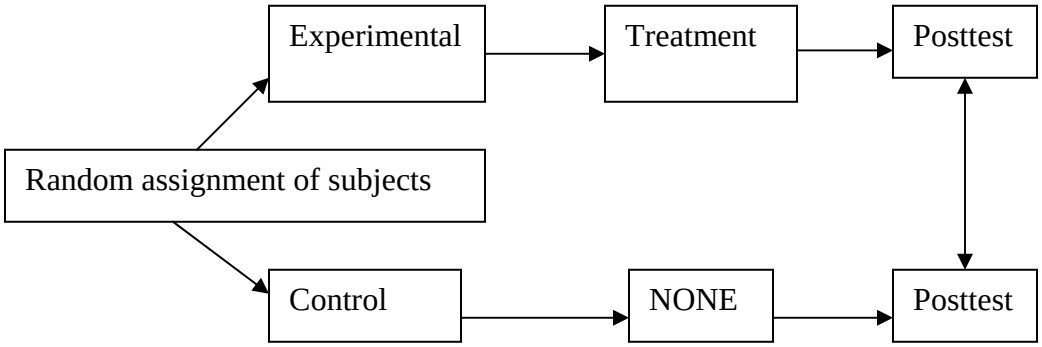

Figure 6: Experiment Design: Random selection and Post-test only

classes. Randomization, identification of experimental and control groups could create a labeling process that might be rejected by parents on the grounds of unequal treatment. Individual teachers and/or administrators for the purpose of a dissertation might conduct a treatment and compare with a regular class. Gains are nearly always demonstrated because of the special efforts of the experimenter. Once the dissertation is completed such teachers or administrators, except a few, rarely continue the strategy. The method is hard to sustain on a widespread basis, mainly because of the impediments within the school as a bureaucracy and the need for schools and teachers to respond to a standardized curriculum and testing.

How the contextual variables function

The question arises as to which design is more efficient and effective in terms of NCLB. In order to make this judgment, we need to develop a holistic perspective of the role of testing in relation to schooling outcomes and students' economic statuses in the industrial society. Testing for selection in practice is the dominant functional aspect of schooling rather than education of all students. NCLB expects the latter but utilizes the former in the form of standardized testing to gain the latter. Based on the findings of Rothstein (203) and Hart and Risley (1995), it is possible that the more NCLB emphasizes standardized testing the less poor people's children might learn as well as middle class students. The design of the occupational structure for industrial productivity determines the educational structure, and subsequently the social structure thereby contributing to self-generated recycling of the system. The industrial occupational structure is based on a hierarchical division of labor requiring training on a stratified basis in knowledge, skills and dispositions. Therefore, the schools are designed from pre-k to grade 12 for selection into the occupational structure. The hierarchical division of labor in the industrial structure is the basis for hierarchical distribution of incomes based on educational levels. Hierarchical distribution of incomes in correspondence with hierarchical education contributes to a hierarchical social structure. Whatever social differences in terms of occupations exist in the larger society are mirrored in each classroom, reproduced and transmitted almost on a

conveyer belt from kindergarten to college. Coleman et al (1965) suggest that it is not the variation in test scores between schools that needs to be explained, but why the test scores tend to vary in the same classroom with the same teacher, same resources and facilities, same curriculum and teacher methodology. These relationships are shown in the following diagram (figure 6) in order to provide an analysis in relation to the issue raised by Coleman, et al. First, all students: lower, middle and high SES (socio-economic status) students enter the classroom in kindergarten and are treated homogeneously in a process of age cohort handling to grade 12. Second, they are administered the same standardized curriculum, teaching methods, tests and rules about discipline as shown in the diagram. The high SES students with an advantage in parental involvement, middle class home curriculum and school appropriate high verbal ability (Bernstein, 1961; Hess and Shipman, 1965; Hart & Risley, 1995) obtain high achievement scores as compared to low SES ones. The variation in test scores occurs also as a result of teachers grading of students on the bell curve (Walberg, 1979). Third, based on differential students' performances, teachers exhibit differential behaviors (Rosenthal & Jacobsen). Teachers' respond to high SES students' achievement by providing higher praise, action

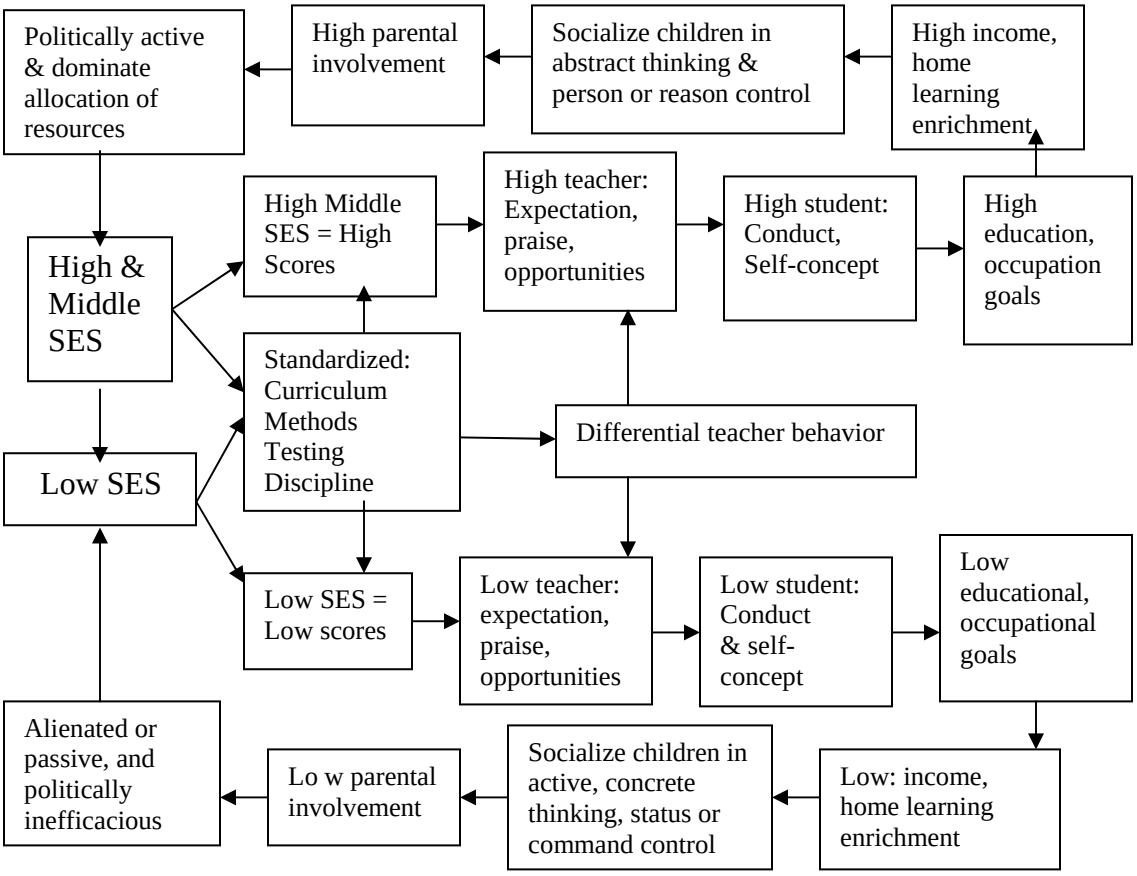

Figure 7: Social structural reproduction through standardization

opportunities and expectation as compared to lower SES students (Good & McCaslin, 1992). High SES students respond with better conduct and higher self-concept, and educational goals that earn high incomes. They obtain higher income jobs, socialize their children in abstract thinking, and are highly involved in school (Hess & Shipman, 1965; Hart & Risely, 1996; Rosenstein. 2003). They are also politically efficacious in the political allocation of resources. Conversely, low SES students obtain lower test scores receive less praise and action opportunities from teachers. They respond with disruptive behaviors and obtain in turn low conduct grades. Consequently, they develop low self-esteem, low expectation about higher education, and occupations, and, as parents, also socialize their children into using language not acceptable in school (Bernstein, 1961; Hess & Shipman, 1965). They are less involved in their children's education at school and are politically inefficacious in the allocation of resources. To break this vicious cycle, it might be needed as Rosenstein (2003) has suggested the development of social and economic policy to treat the socio-economic conditions, directly. Correspondingly, an evaluation approach must consider the social context variables for explaining students' outcomes to facilitate this judgment, and the evaluation methodology ought to include correlation analyses.

The need to research the context in upfront and backend evaluation

We have argued that a major function of evaluation is to examine the extent to which the plan has resolved the problem it was designed to solve in a given context. Obviously, if the planning process did not define the problem accurately, the plan would fail. The question is whether evaluation should just demonstrate that the plan was ineffective, or should it also evaluate the extent to which the plan defined the problem in the context accurately. If outcome evaluation is utilized in conjunction with various forms of experimental designs the contextual variables would not be measured directly and we would not be able to determine whether teacher planning, and teaching and assessment are relevant to specific contributory contextual variable(s). Further, policy-makers, planners and teachers would continue to ignore the contextual variables. It might well be that only the incremental research by each teacher in each classroom followed by appropriate teaching designs based on the research of the contextual variables in each classroom might raise our consciousness about the enormity of the socio-economic hurdle and, therefore the need for policy change as suggested by Rosenstein (2003).

In the meanwhile, each principal could facilitate the evaluation process in contributing to all students to learn over time by providing evaluation leadership in organizing evaluation at grade levels to be monitored by each grade chair of the grade achievement team (GAT) and coordinated at the school level by a school achievement team as shown in the following diagram (figure 8). The principal could organize grade chairs to represent each grade level on a school achievement team (SAT). Each grade chair could be trained as trainers in evaluation, assessment and research to plan. Each grade chair could train grade level teachers in the process. Each teacher should research the conditions of the learners that contribute to variance in performance and to plan and teach for change process. If the teacher was not successful, they could revise their lesson plans and methodologies, etc., accordingly. In addition, student counseling, special programs, after-school programs, staff development, school facilities planning and parental involvement could all be developed in

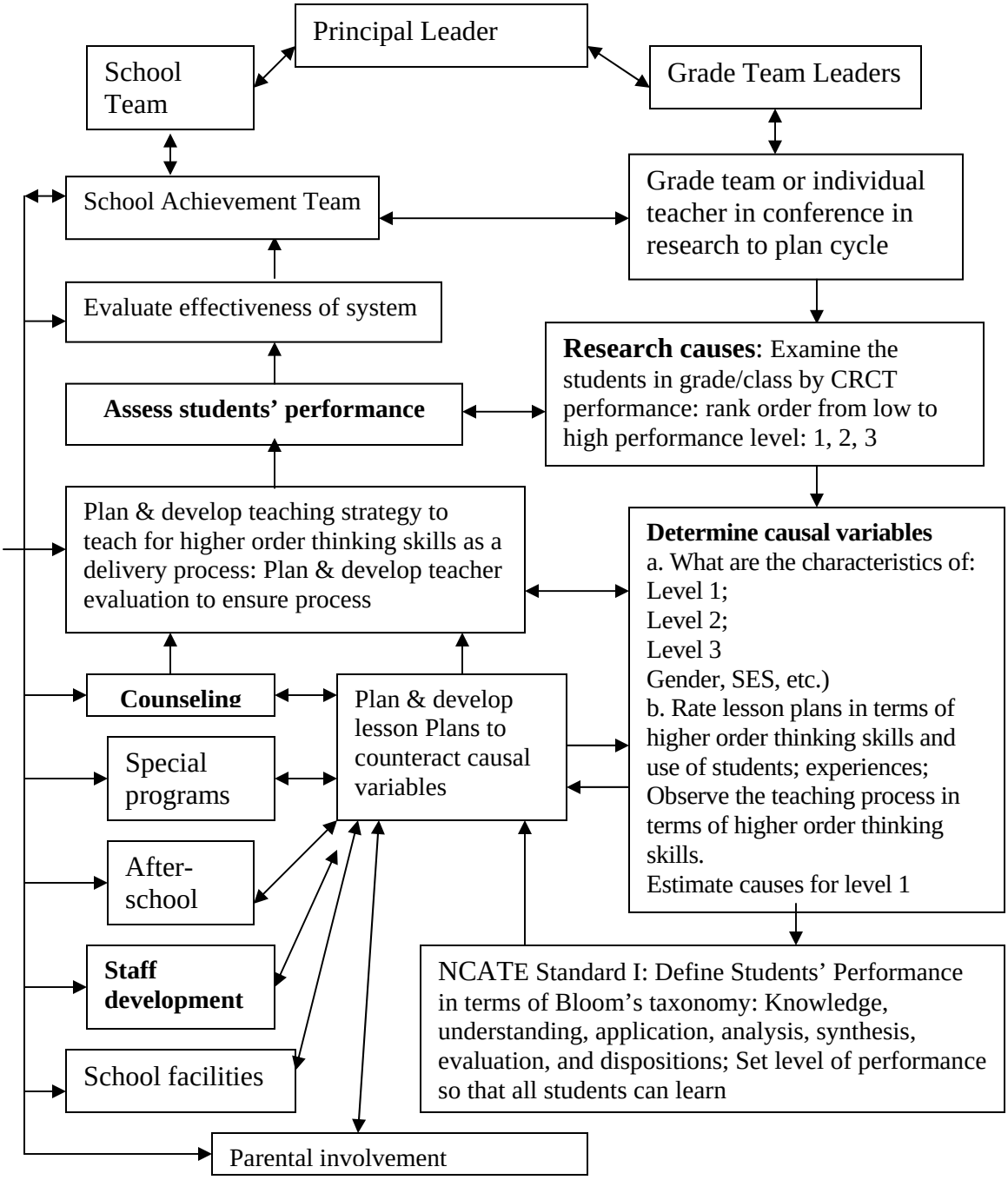

Figure 8: Leader-follower in High definition planning cycle for planning system

alignment to the social contextual variables with the collaboration of the grade team and school achievement team. Each additional strategy could be evaluated with respect to coping with the specified contextual variables in relation to contributing to student achievement. This holistic approach could enhance students' achievement for all.

If the context is not to be measured directly, then:

1. The Federal State Department, and state departments should demonstrate in experimental and correlation studies that all teachers teaching the same standards and utilizing the same methods, etc. would equalize students' performance given the baseline differences in performance before requiring school districts, schools and teachers in classrooms to practice the implementation of the same standards by scripted curriculum and methodologies.
2. The private sector must demonstrate in true-experimental design the "best practices programs" to be marketed to schools significantly enabled all students to perform at proficiency level on state standardized tests. One would assume that the private sector would do as the Pharmaceuticals are doing with respect to drugs being submitted to the USDA for approval. The State department should set up an oversight committee to ensure that the program plan would result in equalizing performance of all ethnic and social class groups before granting approval for such a program plan.

References

Bernstein, Basil (1961). Social class and linguistic development: A Theory of Social Learning. In A.H. Halsey, Jean Floud, & C.A. Anderson (Eds.), Education, Economy, and Society. Glenco, Ill.: Free Press.

Bloom, B. S. (1956). Taxonomy of educational objectives. Handbook I: Cognitive domain. New York: David McKay.

Bloom, B. S. (1976). Human characteristics and school learning. New York: McGraw-Hill.

Cartwright, T.J. (1973). Problems, solutions, and strategies: A contribution to the theory and practice of planning. Journal of American Institute of Planning. May, pp. 179-187.

Chen, H. T. (2006). Practical program evaluation. Thousand Oaks: Sage.

Clark, R. M. (1983). Family life and school achievement: why black children succeed or fail. Chicago: University of Chicago Press.

Coleman, J. S., Campbell, E.G., Hobson, C.J., McPartland, J., Mood, A.M., Weinfield, F.D., & York, R.L. (1966). Equality of educational opportunity. Washington, D.C: U.S. Department of Health, Education, and Welfare. Government Printing Office.

Cronbach , L. J. et al (1980). Toward reform of program evaluation. San Francisco: Jossey-Bass.

Cunningham, W. G. (1982). Systematic planning for educational change. Palo Alto, CA: Mayfield Publishing Company.

Curwin, R. L. & Mendler, A. N. (1988). Packaged discipline programs: Let the buyer beware. Educational Leadership. 46(6), 68-71.

Deming, E. W. (1986). Out of the crisis. Cambridge, Mass: Massachusetts Institute of Technology Center for Advanced Engineering.

Dewey, J. (1916). Democracy and Education. New York: MacMillan.

Fetterman, D. M. (2000). Foundations Of empowerment evaluation,. Thousand Oaks, CA: Sage.

Fitzpatrick, J. L., Sanders, J. R. & Worthen, B. R. (2004). Program Evaluation: Alternative Approaches and Practical Guidelines. Boston, MA: Pearson Publications Inc.

Drucker, P. F.(1964). Management for results, New York: Harper & Row

Flanders, N. A. (1970). Analyzing teacher behavior. Reading, Mass.: Addison-Wesley.

Freire, P. (1970). Pedagogy of the oppressed. New York: Seabury Press.

Freire, P. (1973). Education as Critical Consciousness. New York: Seabury Press.

Fullan, M., & Miles, M. (1991) Getting educational reform right: What works and what doesn't. Unpublished paper. University of Toronto: Center for Policy Research.

Good, T.L., & McCaslin, M. M. (1992). Teaching effectiveness. In Marvin C. Alkin (ed.), Encyclopedia of educational research, 6th ed., Volume 4, pp. 1373-1388. New York, McMillan Publishing Co.

Glass, G. V., McGraw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills & London: Sage Publications.

Gredler, M. E. (1996). Program Evaluation. Pearson Education: NJ: Prentice-Hall. Inc.

Hart, B. & Risley, T. R. (1995). Meaningful differences in the everyday experiences of young American children. Baltimore: Paul H. Brooks Publishing Company.

Hess, Robert D, & Shipman, Virginia C. (1965). Child Development. Vol. 36, No. 4, December. pp. 869-85.

House, E. R. (1980). Evaluating with validity. Beverly Hills, CA: Sage.
Hossler, D. & Stage, F. K. (1992). Family and high school experience influence on the postsecondary educational plans of ninth-graders, American Educational Research Journal, 29, 425-451.
Ishikawa, K. (1976). Guide to quality control. Asian Productivity Organization, pp. 24-26.
Kerlinger, F. N. (1992). Foundations of behavioral research, (3rd. ed.). Fortworth, TX: Harcourt Brace.
Leinhardt, Gaea (1992). What research on learning tells us about teaching, Educational Leadership. April, pp. 20-25.
Levine, Arthur (2006). Educating School Teachers: The Education Scholars Project. Washington, DC 2—36.
Mescon, M. H., M., Albert, & Khedouri, F. (1988). Management (Third edition), New York: Harper & Row
Metfessel, N. S. & Michael, W. B. (1967). A paradigm involving multiple criterion measures for the correlation of the effectiveness of school programs. Educational and Psychological Measurement, 27, 931-943.
Mc Gregor, D. (1960). The human side of enterprise. New York: McGraw-Hill.
Odiorne , G. G. (1965). Management by Objectives. Belmont, California: Pitman Learning.
Patton, M. (1990). Qualitative evaluation and research method (2nd ed.). Newbury Park: Sage.
Olson. L. (2006, January 5). A decade of effort. Education Week, 25(17), Pp. 9-10.
Patton, M. Q. (1997). Developmental evaluation. EvaluationPractice., 15, 311-320.
Persaud, G., & Turner, T. (2002). High definition planning for effective schools: reflections from the field. *Educational Planning, 1*4(1), 65-76.
Persaud-White, T. (1996). High definition planning and teaching for student academic achievement. Paper presented at the Georgia Educational Research Association, November.
Persaud, G., & Turner, T. (2002). High definition planning for effective schools: reflections from the field. Educational Planning, 14(1), 65-76.
Persaud, G., & Turner, T., & Moffett, N. (2007). Conducting Teacher Evaluation to Improve the Teaching of Higher Order Thinking Skills: Use of an Observation Based Instructional Assessment (OBIA). National Evaluation Institute Conference: Florida, October 4-7, 20; Retrievable from Google: Search: Ganga Persaud-Locate NEI conference
Provus, M. (1971). _Discrepancy Evaluation. Beverly Hills, CA: McCutchan
Rothstein, R. (2004). Class and schools: Using social, economic and educational reform to close the Black-White achievement gap, Washington, D.C. Economic Policy Institute; also distributed by, New York: Teachers College Press
Sanders, J.R. (ed., Education Evaluation: Theory and Practice. Worthington, Ohio: Charles A. Jones Publishing Co. 1973, p.139.
Stake, R. E. (1967). The countenance of educational evaluation. Teachers College Record, 68, p. 623-540.
Stufflebeam, D. L.(1973). Education Evaluation and Design. In Worthen, B. R. and

Stufflebeam, D. L. (1973). Education Evaluation and Design. In Worthen, B. R. and Sanders, J.R. (ed., Education Evaluation: Theory and Practice. Worthington, Ohio: Charles A. Jones Publishing Co. 1973, p.139.

Tough, Paul (2006). What it takes to make a student. New York Times, November 26 Read article at: http:/www.mytimes.com/2006/magazine/26tough.html?

Tyler, R. (1940). Basic principles if curriculum and instruction. Chicago: University of Chicago Press.

Walberg, H.J. (1979) Educational environments and effects: Evaluation, policy, and productivity. Berkeley, Cal.: McCutchan Publishing Corporation.

Worthen, B. R. and Sanders, J.R. (1973). Education Evaluation: Theory and Practice. Worthington, Ohio: Charles A. Jones Publishing Co. 1973, p.139.