Evaluating the Validity of State Accountability Systems

Scott Marion and Brian Gong
Center for Assessment

The 2003 Reidy Interactive Lecture Series

Nashua, NH October 9, 2003

A little exercise (part 1)

- On a piece of paper, write down the percentage of schools in your state that you consider low performing and in need of outside intervention.
- Feel free to indicate your state, but you don't have to.

A little exercise (part 2)

- For those of you with children, grandchildren, nieces, nephews, etc., please indicate the following:

1. The % of schools where you'd be HAPPY to send your kids?
2. The % of schools where you'd be WILLING to send your kids?

"Validity of accountability systems"—can you answer: (1)

- Why did your school get a different rating under NCLB than it did under the state's previous system?
- Why is your state's percentage of schools identified different than another state's?
- What evidence do you have that the improvement goals are reasonable?

"Validity of accountability systems"—can you answer: (2)

- How possible is it that one child made the difference between a "good" and a "bad" school rating? How did you protect against an unfair "good class/bad class bounce"?
- How are you sure that people didn't cheat on reporting dropout data, mobility, LEP classification, or any other data?

"Validity"—can you answer: (3)

- How could people "game" the system, and what steps have you taken to detect and prevent that?
- How do you account for performance differences—between different parts of the state, subgroups, or content areas—and what are you going to do about it?
- What evidence do you have that that will help?

Validity – A crucial concern

- Validity is central to having a credible and fair accountability system. It is the most important technical criterion for defending the quality of the accountability system.
- The USDE NCLB workbooks require "validity and reliability" evidence (plan)
- The framework and research agenda being presented today are designed to assist states to evaluate and defend the validity of their accountability systems.

The Center's Accountability System Validity Framework

I. The Theory of Action (How do you think things should work?)

II. Accuracy and Consistency of School Classifications (Did you identify the right schools?)

III. Consequences of the System (Did people do what was expected? what was fair? what was effective?)

IV. Interventions (Did people—including the state—do what was legally required and what was needed to help students, schools, and districts succeed?)

A Theory of Action

A Theory of Action

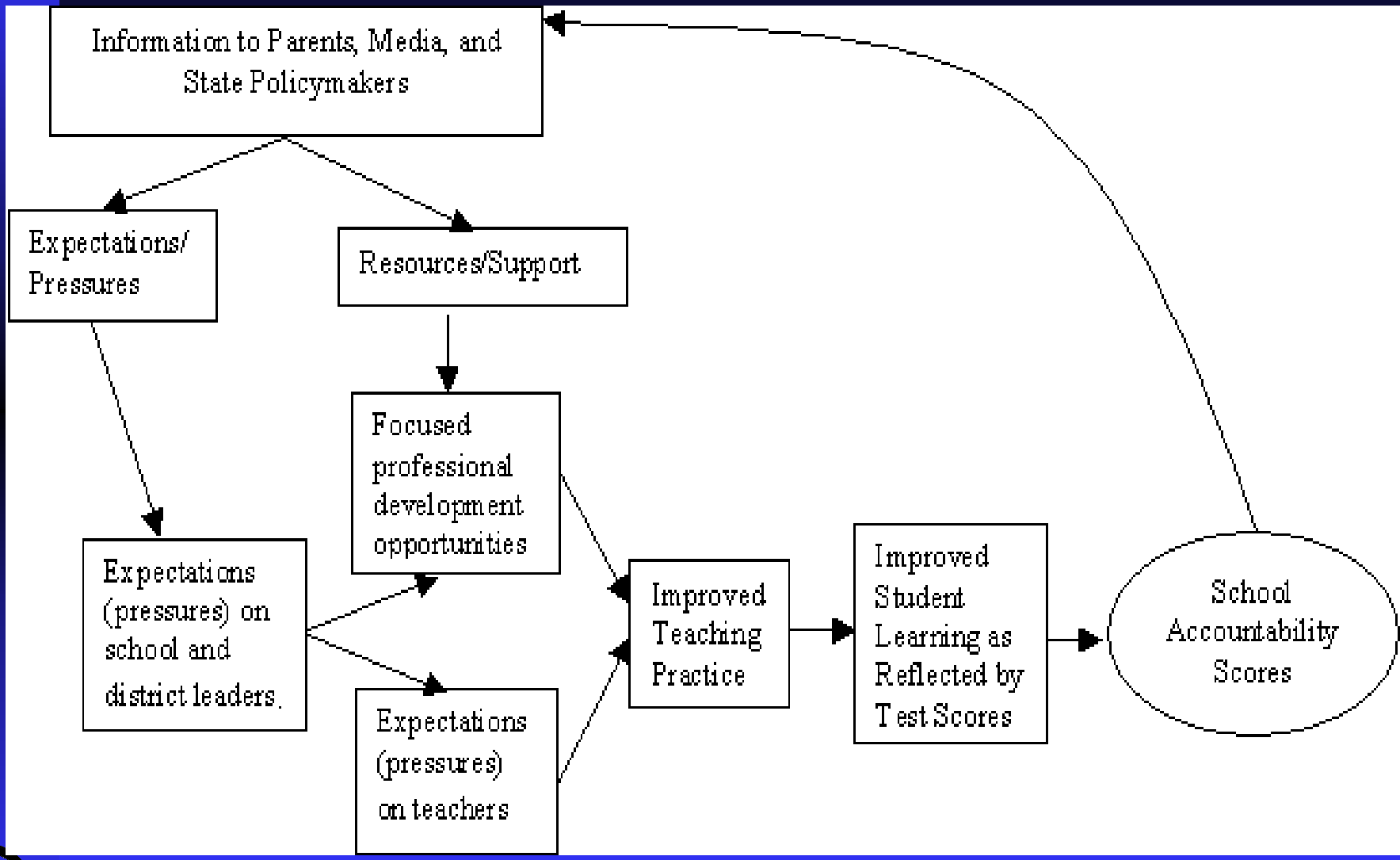
Similar to the evaluation of assessment validity, where purposes and uses must be specified, the evaluation of accountability system validity must specify **how and why** the system is intended to work in order to **improve student learning and system capacity**.

Cronbach referred to a "nomonological net" as a way to specify how the theory and actions should interact to explain the score inferences.

Making values explicit

- The theory of action is necessarily a values statement. NCLB is a values statement and the State should clarify the alignment of its values with those of NCLB.
- State leaders and stakeholders need to come to shared understandings of such things as the definition of a good school (as specifically as possible), appropriate actions, and what constitutes a valid accountability system.
- This values statement will serve as a touchstone during the data collection, analysis, and evaluative phases of the study.

A Theory of Action Example



Expected and Intended Actions

- By testing students in the requisite grades and rewarding and sanctioning schools, the expectation is that low-performing students will learn more and that all subgroups will be helped to close any gaps.
- That's a big jump (or leap). What do you think are some of the intermediate steps? Take a few minutes and draw a picture of the intermediate steps.

Integrating Values and Actions

- Making sure the intended actions reflect the state's values about teaching and learning, motivation, educational change, etc. is a challenging and time-intensive aspect of developing a state accountability system.
- Ideally this is done explicitly in the design phase. If not, the evaluator should attempt to infer the values from the design and through interviews and analysis.

Accuracy and Consistency of School Classifications

Accuracy and Consistency

Are the "right" schools being identified and not identified?

- Is your design aligned with your values?
- Do your technical procedures & policies implement your design appropriately?
- Do certain factors, such as grade span, additional indicators, and the number of subgroups, influence the accuracy of your design?

Accuracy and Consistency

- Consistency is a necessary, but not sufficient condition for ensuring the validity of large-scale accountability systems.
  - What is the effect of sampling error on decision consistency?
  - What is the influence of multiple conjunctive tests on decision consistency?
  - What are the effects of error other than sampling error on decision consistency?

School scores as a sample…

School data should be treated as a **sample** from a larger population…. *To conclude on the basis of an assessment that a* ***school is effective as an institution*** *requires the assumption…that the positive outcome would appear with a student body other than the present one, drawn from the same* ***population*** (Cronbach, Linn, Brennan, & Haertel, 1997, p. 393, emphasis added).

Strategies for dealing with the effects of sampling error

- Raising the minimum-n
- Using confidence intervals
- Adjusting for multiple conjunctive "hurdles"

These issues have been addressed in previous RILS and other publications, e.g.:

Hill, R. K. & DePascale, C. A. (2003). Reliability of No Child Left Behind Accountability Designs. Educational Measurement: Issues and Practices

A Quick Minimum-n Note

- In reality, raising the minimum-n to levels high enough to have a noticeable effect on reliability—especially when applied to "safe harbor" improvement conditions—would require such large samples that would be impractical for many states. It becomes a consequential validity issue!

Consequences of Minimum-n

- Raising or lowering the size of the minimum-n is more of an issue of consequential validity than an issue of reliability. Raising the minimum N simply allows small schools and subgroups an easier way through the system.
- What if the minimum-n was 30 and you had a school with a subgroup of 25 that consistently had fewer than 10% proficient (with a target, for example, of 40% proficient and increasing) for that group? Wouldn't you be confident in saying that this subgroup is not being educated appropriately?

Accuracy and Consistency:
Balancing Type I and II Errors

- **Type I Error** in an accountability context means identifying a school as failing AYP if it did not "truly" fail ("false positive").
- **Type II Error** means that a school that should have been identified as failing AYP was incorrectly considered to be a "passing" school ("false negative").

Type I and II Errors

- We do not question the importance of minimizing Type I errors, given the sanctions associated with NCLB including the effects of multiple conjunctive decisions on the nominal Type I error rate.
- Yet, few states appear concerned with minimizing Type II errors.
- We think it is important to identify schools for improvement if students are not being served adequately. (See R. Hill 2003 RILS presentation for a "two-tier" approach for addressing both types of errors.)

Examples of Studies

- Consistency
  - Numerous examples of consistency studies have been laid out by Hill and DePascale: www.nciea.org
- Validity
  - Many possibilities. We present a few, tied to aspects of the framework, on subsequent slides.

Accuracy-1

- You can do a simple correlational study to examine the relationship between the number of subgroups and the likelihood of a school being identified.
- Is the validity threatened if schools' chances of being identified is simply a function of the presence of subgroups?

Accuracy-2

- A more appropriate and only somewhat more complex approach would be to use logistical regression to study the influence of specific subgroups and/or combinations of subgroups on the probability of a school being identified.
- How would the validity of the system be affected if we found that the presence of one or two specific subgroups significantly influenced a school's likelihood of being identified?

Type I & Type II Errors (CI)

- Gather additional data about schools that "made AYP" due to CI, such as historic achievement data, school climate information, etc.
- Design a systematic approach (qualitative or quantitative) for evaluating, based on the full set of data, whether or not these schools should "truly" be considered passing—TIED TO YOUR VALUES.
  - "Truly" passing—protected against Type I.
  - Should have failed=Type II error.
- Calculate the percentage of schools correctly classified as a result of using confidence intervals.
- Similar analyses can also be done using discriminant analyses procedures.

Type I and Type II (min-n)

- Select a sample of schools that failed to make AYP because one or more subgroups missed the AMO by 5% proficient or less, but no subgroup failed by more than this amount.
- Gather additional data about the school.
- Design a systematic approach for evaluating, based on the full set of data, whether or not these schools should "truly" be considered failing.
  - Incorrectly labeled failing=Type I error
  - "Truly" failing—protected against Type II error
- Calculate the percentage of schools correctly classified as a result of using this minimum-n.

Type I and Type II (min-n)

- Establish a relatively low (e.g., n=5) minimum-n.
- Compute the % proficient for each subgroup for each school.
- For each subgroup, divide the schools into two groups—those below the accountability minimum-n (e.g., n=30) and those above the accountability minimum-n .
- Estimate the population distribution of school means for each of these two groups of schools.
- If the two population means are not significantly different, we can conclude that schools are not identified simply because they fall below min-n, a Type II error.

Results of one set of analyses

- Is the distribution of student performance different for low SES students in schools that meet the n=30 rule compared with schools where 6<n<30?
- 4th grade: n>30 = 10 schools; 6<n<30 = 114 schools. The mean % proficient was significantly lower in schools that met the n=30 criterion compared with schools that did not for both ELA and math ($p<0.001$;$p<0.01$).
- 8th grade: n>30 = 19 schools; 6<n<30 = 35 schools. No statistical difference between the two groups for both ELA and math.
- What do these findings tell us (if anything)?

Consequences of the Accountability System

Types of Policy Consequences

Intended Positive	Unintended Positive
Intended Negative	Unintended Negative

- Cells on the shaded diagonal are the focus. We're willing to suspend conspiracy theories regarding the "intended negative" and willing to take the "unintended positive" as a delightful, but rare occurrence.

Intended Positive

- "…to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments."
- We, as evaluators, should look to see whether or not these outcomes have been achieved.
- What are some questions would we ask to address this issue?

Unintended Negative

- Theorists (e.g., Cronbach, Stake, House, Shepard) have made a convincing case that evaluators must search for and make explicit the unintended negative consequences caused by programs or policies.

- Take a few minutes to generate a list of positive INTENDED and negative UNINTENDED consequences that you've either seen or are concerned about.

A Categorization of Potential Consequences:

A. Effects on Title I programs, budgets, personnel, and students
B. Definitional (design) Effects
   - FAY
   - SWD and ELL placements
C. Closing the achievement gap

A study example

What effect does FAY (full academic year) have on school/district identification?

- Who is out of accountability because of FAY—at the school, district, state levels?
- What are school/district characteristics with high exclusion because of FAY?
- What happens to mobile students instructionally? Is district accountability "working" for these FAY students?

Another study example

- Gather data about the qualifications of leaders and teachers in Title I and non-Title I schools.
- Monitor the trends in teacher and leader quality to see if there is a relative loss in educator quality in Title I schools.

Interventions To Improve Teaching and Learning

Coherent and comprehensive accountability system

- How does your "Theory of Action" reflect your "Theory of Assistance"?
- What are your intended positive and intended negative consequences in terms of interventions to improve teaching and learning?
- What unintended negative consequences are you protecting against?

Assistance and "accountability system validity"

- Evaluation is focused on the "right" things
- Performance of schools is accurately portrayed
- Information (goals, targets, performance, consequences) is used appropriately in accountability
- Appropriate schools are identified to receive consequences
- Schools/districts receive appropriate consequences
- AND THEN… [black box, local control, individual implementation…]
- Students and schools are "better" than they would be without the accountability system

NCLB consequences delineated

- NCLB provides clear delineation of some consequences for schools, districts, and schools
- What set of consequences is chosen by the state?
- What other "actions" are planned by the state, district, school, others?

NCLB School Consequences (by LEA)

- Year 1 : school identified as failing AYP once; (Title 1 schools identified under old ESEA treated as Year 2)
- Year 2: (school identified as failing AYP twice); public school choice, supplemental services, improvement plan
- Year 3: choice, supplemental services, technical assistance from district
- Year 4: Yr. 3 plus at least one corrective action (a) replace staff; b) new curriculum; c) decrease management authority at school; d) outside expert; e) extend school year or day; f) restructure internal school organization
- Year 5: choice, supplemental services, and plan for alternate governance (charter; replace staff; contract management; state take-over, other)
- Year 6: implement alternate governance plan

NCLB District Assistance (by state)

- Year 1 (during ’02-’03): Technical assistance from state, if requested by district
- Year 2: [district identified by state as failing AYP once]
- Year 3: [district fails AYP twice] Corrective action by state (at least one): a) reduce funding; b) new curriculum; c) replace staff; d) remove school from district governance; e) appoint receiver; f) abolish/ restructure district; g) authorize and pay for inter-district student transfer) [if g), must do at least one additional]
- Year 4: Any additional sanctions from Year 3 list, as deemed appropriate by state

NCLB – Assistance to Students

- Ongoing
  - Title I services
  - Curriculum/instruction grant programs
- If school identified as "In Need of Improvement"
  - Supplemental services
  - Public school choice, transportation
  - (District/state technical assistance)

NCLB – Possible school assistance

- Technical assistance (e.g., Needs Analysis)
- Monitor planning and other processes
- Funding (?)
- Direct services of "change agents," C&I specialists
- Change leadership, governance structures

Intervention/consequences validity

- You may ask what evidence supports the NCLB intervention specifications
- You should be prepared to strongly defend your intervention plan
  - What is your plan, including practical constraints (e.g., “triage”)?
  - What evidence do you have that it works and is fair?
  - What will you do to improve your system?

Action paths to learning and system capacity

- Changes occurred because…
  - Inappropriate score changes
  - Motivation, effort, focus, leadership [do same thing better]
  - Changes in curriculum
  - Changes in instruction
  - Changes in inclusion
  - Teacher transfer
  - Student transfer
  - Systemic changes
  - ?
- How can this be done in another school…

Sample study

- How effective was school choice?
  - What school choice was offered? How? Who transferred? Why? What happened? Why?
  - What happened to students who transferred? – programs, services
  - What happened to students who did not transfer?
  - What school rating changes are due mostly to population shifts?

More study topics

- How differentiated is your assistance in terms of NCLB "hurdles"
- Do you differentiate in terms of different contexts (e.g., small/rural schools)?
- Do you identify and account for differences in other resources (e.g., teacher "quality)?
- Do you track the quality of implementation?

The Center's Accountability System Validity Framework

I. The Theory of Action (How do you think things should work?)

II. Accuracy and Consistency of School Classifications (Did you identify the right schools?)

III. Consequences of the System (Did people do what was expected? what was fair? what was effective?)

IV. Interventions (Did people—including the state—do what was legally required and what was needed to help students, schools, and districts succeed?)