# Common Problems with Accountability Systems

Richard Hill

The Center for Assessment (NCIEA)

Presentation at the

*Conference on Policy and Measurement Issues*

*in Large-scale Science and Mathematics Assessment*

Washington, DC, March 23-25, 2000

Sponsored by RAND and NSF

# Four Major Issues

- Validity
- Consistency of metric
- Consistency of included population
- Reliability

- Importance of these issues for accountability vs. merely reporting

# Validity

- Validity of test results
  - Content related to frameworks
  - Coaching/Cheating
- Validity of accountability system
  - Does it provide incentives for the actions you want people to take
  - Does it provide disincentives for the actions you don’t want people to take
- Breadth of measures
- Realistic goals for improvement

# Consistency of Metric

- Consistency of content
- Equating
- Consistency of administration
- Consistency of motivation
- Consistency of standards across time and across grades
- Consistency of scoring

# Consistency of Included Population

- In a system with performance levels, students don’t have to be tested to be included in accountability
- Most robust systems include all students

# Reliability

- Two ways to estimate
  - Split half
  - Estimation of variance components
- Needs to be done on actual population, since conditional probabilities don’t tell entire story

# Reliability

- Classification error is appropriate way to determine accuracy
- Error rates are higher than most people think
  - Split half study identified 18 schools with at least one half below standard

# Reliability

- Problem is *not* measurement error, but sampling error
- Error is reduced, but only somewhat, by following cohorts or including adjacent grades
  - Students move
    - Lose from accountability system, or
    - Lose advantage of tracking cohort
  - r = .70

# Reliability

- More difficult to estimate gain than performance
  - True variance smaller
  - Two samples rather than one
  - Split half study
    - r = .96 for performance
    - r = .70 for gain from one year to next

# Common Errors

- Conjunctive decision rules
- Coarse reporting statistics
- Too short a waiting period
- Identifying extreme cases

# Conjunctive Decision Rules

- Example:
  - two identical schools of 200 students each
  - one has two subgroups of 100 each
  - Each school has a Growth Target of 15 points, and a standard error of 15 points

# Conjunctive Decision Rules

- If each school improves by 15 points:
  - School A has 50 percent probability of succeeding
  - School B has a 12.5 percent probability of succeeding

    (.5 * .5 * .5)

# Conjunctive Decision Rules

- If each school improves by 30 points:
  - School A has 74 percent probability of succeeding
  - School B has a 35 percent probability of succeeding

$(.74 * .69 * .69)$

# Coarse Reporting Statistics

- Split-half analysis, between 40 and 71 students in each half

| | |
|---|---|
| SS: | r = .92 |
| Index of 1-5: | r = .89 |
| Index of 1-4: | r = .87 |
| Pass/Fail: | r = .84 |

# Coarse Reporting Statistics

- SS: $(1-r^2) = .15$
- Index of 1-5: $(1-r^2) = .21$
- Index of 1-4: $(1-r^2) = .24$
- Pass/Fail: $(1-r^2) = .29$

- Earlier example of 1/16—revised procedures gave 18/35

# Too Short a Waiting Period

- Two groups
  - Each has 100 schools
  - Each school has 100 students
  - Each starts at state average
  - Each has to improve 1/20 of distance to long term target
  - Standard error = Growth Target
- Group A actually improves
- Group B makes no change

# Too Short a Waiting Period

| Change in Score from Year 1 to Year 2 | Group A (Actually Improved) | Group B (No Real Improvement) | Total |
|---|---|---|---|
| Gains greater than or equal to Growth Target | 50 | 35 | 85 |
| Improvement, but not as much as GT | 15 | 15 | 30 |
| Decline | 35 | 50 | 85 |
| Total | 100 | 100 | 200 |

# Too Short a Waiting Period

- Average two years
- Improvement in two areas
  - Twice the distance
  - Half the error variance

# Too Short a Waiting Period

| Change in *Average* from Years 1 and 2 to Years 3 and 4 | Group A (Actually Improved) | Group B (No Real Improvement) | Total |
|---|---|---|---|
| Gains greater than or equal to Growth Target | 50 | 12 | 62 |
| Improvement, but not as much as GT | 38 | 38 | 76 |
| Decline | 12 | 50 | 62 |
| Total | 100 | 100 | 200 |

# Identifying Extreme Cases

- Example 1:  Earlier example—even 18/35 a marginal result
- Example 2:  Observing that small schools have greatest increase in scores
- The probability of being classified in top category two consecutive cycles is close to 0

# More Detail

- Second Reidy Interactive Lecture Series, October 5 and 6
- Publication of Lecture proceedings and standards
- Proceedings of first lectures to be available ~ June 1