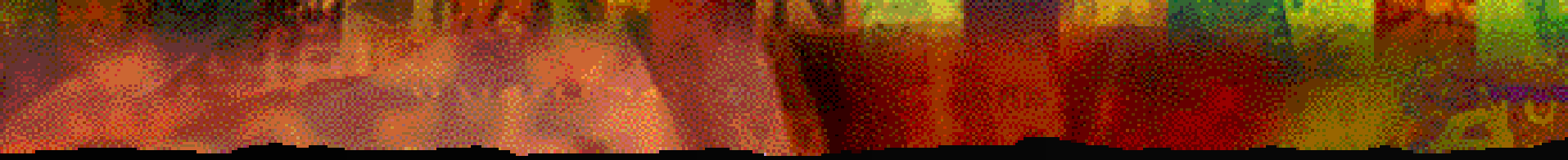


Third Annual Reidy Interactive Lecture Series

Validity of Accountability

Four Major Criteria

1. Validity of Accountability Design—Are we asking the right questions of the data?
2. Validity of Assessment Gains—Do increases in test scores statewide reflect real gains in achievement?
   - Scale
   - Population included in accountability
   - Testing conditions

Four Major Criteria (cont'd)

3. Reliability—Would we make the same decision about a school if we ran another year's data through the same process?
4. Stability—Will unchanged schools get the same designation in successive years?

Four Different Ways of Defining "Quality"

	Status	Change
Achievement		
Effectiveness		

Four Different Ways of Defining "Quality"

	Status	Change
Achievement	How well do students score?	
Effectiveness		

Four Different Ways of Defining "Quality"

	Status	Change
Achievement	How well do students score?	Are this year's scores higher than last year's?
Effectiveness		

Four Different Ways of Defining "Quality"

	Status	Change
Achievement	How well do students score?	Are this year's scores higher than last year's?
Effectiveness	How much do students learn between 3rd and 4th grade?	

Four Different Ways of Defining "Quality"

	Status	Change
Achievement	How well do students score?	Are this year's scores higher than last year's?
Effectiveness	How much do students learn between 3rd and 4th grade?	How much more are students learning between 3rd and 4th grade than they did last year?

Four Different Ways of Defining "Quality"

	Status	Change
Achievement	▪52% Proficient ▪55th %ile	▪5% more Proficient ▪%ile increase of 3
Effectiveness	Change from gr. 3 to gr. 4 = 20 SS points	Increase in change from gr. 3 to gr. 4 = 4 more SS points

Comparing the Schools—Model A

Model	School A		School B		School C		School D	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
A (02)	60.3	4	88.3	1	70.3	3	77.0	2

Comparing the Schools—Model B

Model	School A		School B		School C		School D	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
A (02)	60.3	4	88.3	1	70.3	3	77.0	2
B (02 – 01)	-1.0	4	-0.7	3	2.0	1	-.3	2

Comparing the Schools—Model C

Model	School A		School B		School C		School D	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
A (02)	60.3	4	88.3	1	70.3	3	77.0	2
B (02 – 01)	-1.0	4	-0.7	3	2.0	1	-.3	2
C (4^{th} + 5^{th})	5.0	1	-1.0	4	0.0	3	1.0	2

Comparing the Schools—Model D

Model	School A		School B		School C		School D	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
A (02)	60.3	4	88.3	1	70.3	3	77.0	2
B (02 – 01)	-1.0	4	-0.7	3	2.0	1	-.3	2
C (4^{th} + 5^{th})	5.0	1	-1.0	4	0.0	3	1.0	2
D	-1.5	3	3.0	2	-2.0	4	5.0	1

Correlation among Models, Using End Results

Quadrant	Quadrant		
	B	C	D
A	.27	.17	.08
B		.46	.53
C			.74

Correlation among Models, Using Starting Results

Quadrant	Quadrant		
	B	C	D
A	-.31	.21	-.23
B		-.18	.53
C			-.50

Quadrant A—Achievement status

- Variations
  - Upper bar and lower bar
    - Identify schools with low SES AND poor teaching or high SES AND good teaching
    - Miss schools with low SES and good teaching, and those with high SES and poor teaching
  - Identification of extreme cases
  - Use regression to partial out SES
  - Set bottom bar and raise it over time

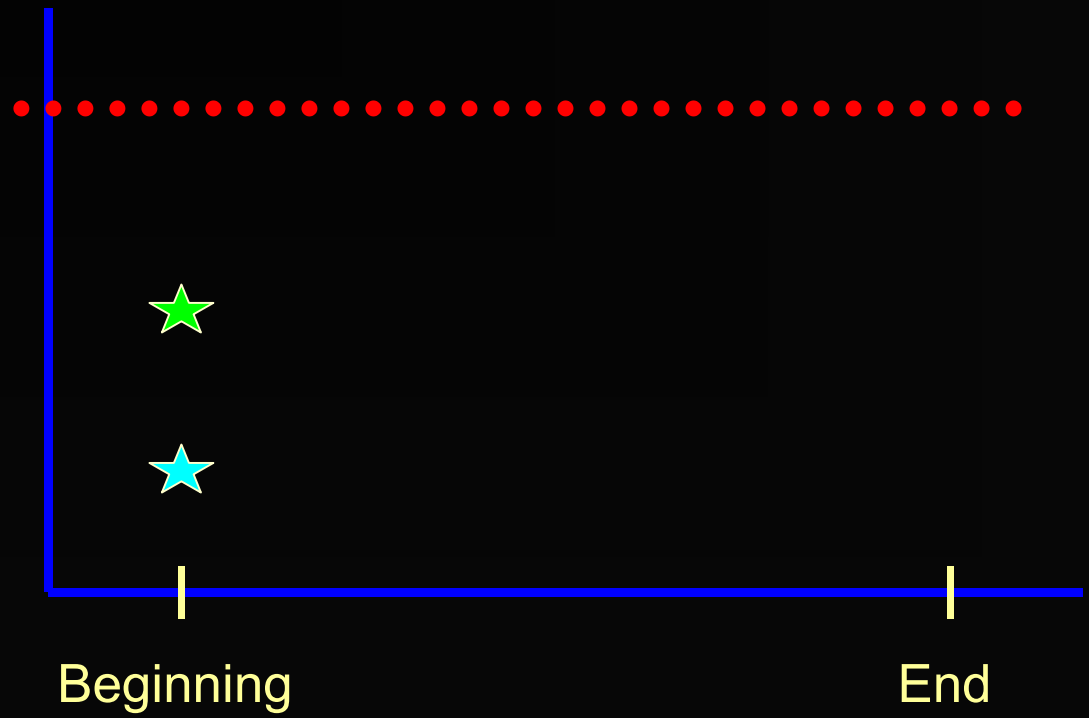
Quadrant A—Achievement Status

- Strengths
  - Reliable
  - Stable
  - Simple to understand
  - Fast to implement
- Assumptions
  - Teachers and schools are completely responsible for student outcomes (2C/D)
  - Low SES students are same challenge as high (3-3I)

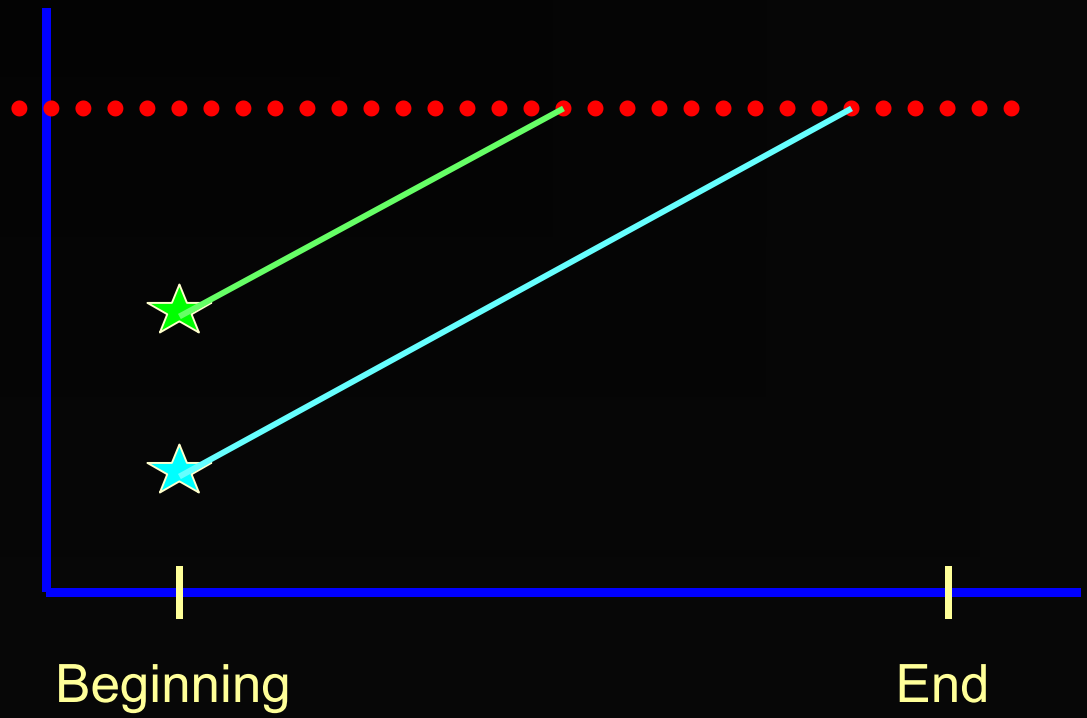
Quadrant B—Achievement Change

- Variations
  - Upper bar
  - Lower bar
  - Improvement expected
    - Same for all schools
    - Time same for all schools

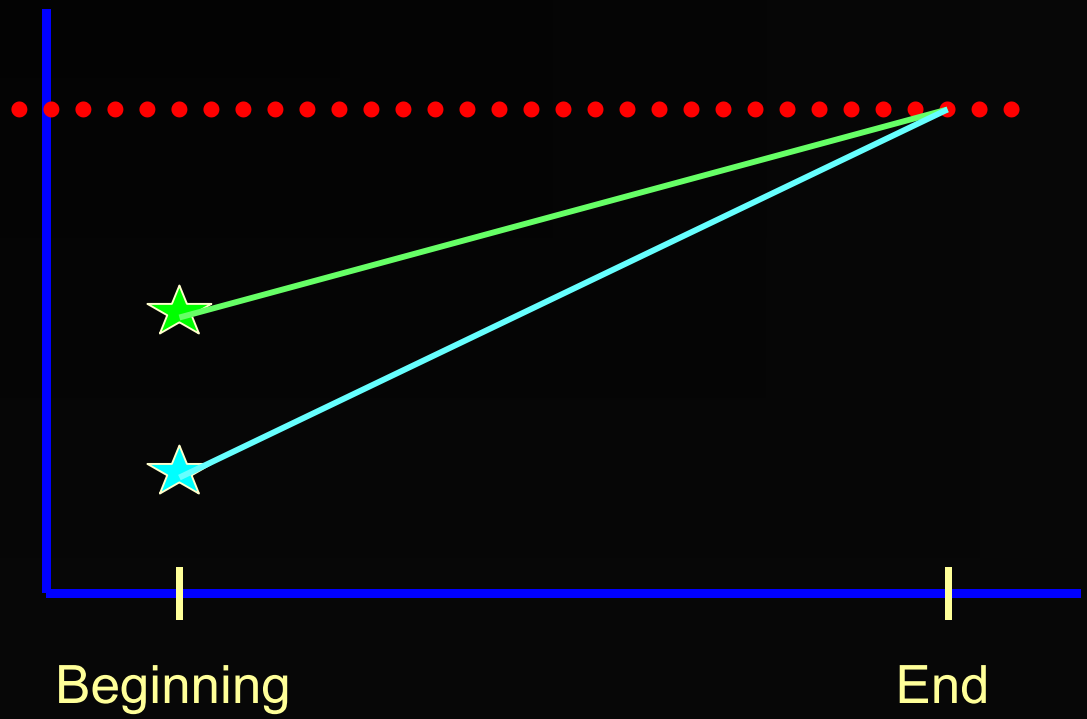
Same Improvement for All Schools



Same Improvement for All Schools



Same Timeline for All Schools



Quadrant B—Achievement Change

- Variations
  - Upper bar (3-3H)
  - Lower bar (3-3G)
  - Improvement expected (3-3J)
    - Same for all schools
    - Time same for all schools

Quadrant B—New Baseline Each Cycle vs. One for Long Period

- New baseline each cycle
  - Negative correlation between consecutive rankings
  - Reliability dependent on amount of gain expected; usually quite low (*cf*. last year's RILS, CA)
- One baseline for long period
  - Importance of accurate baseline
  - Consistency (for long period)
    - Scale
    - Population included in accountability
    - Testing conditions

Quadrant B—Achievement Change

- Strengths
  - Assures upward movement
  - Fairer for low-SES schools than Quadrant A
- Assumption
  - Everyone is expected to improve, regardless of whether they already were strong (3-3G/H)

Quadrant B—Subtle Point

- This model may be far more appropriate for *state* accountability than *school* accountability

Quadrant C—Effectiveness status

- Variations
  - True-longitudinal design (matched students)
  - Quasi-longitudinal design (unmatched cohorts)

Quadrant C—Effectiveness status

- Strengths
  - Closest fit to typical definition of “effective teaching”
- Weaknesses
  - Requires testing of consecutive grades
  - True-longitudinal
    - Ns may be small
    - Requires ability to track students across years
    - Excludes students (disproportionately)
  - Quasi-longitudinal
    - *May* be poorly correlated with TL results
  - Excludes lower grades from accountability
  - Not necessarily any growth over time

Quadrant C—Effectiveness status

Grade	Year			
	1	2	3	4
3	45	41	37	33
4	53	49	45	41
5	61	57	53	49

Quadrant C—Effectiveness status

- Requires pre-test scores (testing at consecutive grades)
- Can be a teacher-level evaluation device
- Analogy to Quadrant A adjusted for SES—only now you're adjusting for pre-test scores rather than SES

Quadrant D—Change in Effectiveness

- Strengths
  - None
- Weaknesses
  - All of Quadrant 2 and Quadrant 3, plus
  - Expected changes are small and hard to detect

Stability Coefficients

Model	Number of Grades		
	1	2	3
A	.85	.93	.95

Stability Coefficients

Model	Number of Grades		
	1	2	3
A	.85	.93	.95
B	-.43	-.32	-.24

Stability Coefficients

Model	Number of Grades		
	1	2	3
A	.85	.93	.95
B	-.43	-.32	-.24
C-TL	.26	.11	-.04
C-QL	.29	.11	-.13

Correlation Between Model B and Model C-QL

Number of Grades of Testing	Correlation
2	.61
3	.78
4	.83

Variations

- Upper bar/lower bar
- Standards
- Amount of time given to meet standards
- Amount of time between accountability decisions
- How expectations for performance or improvement are generated

Variations (cont'd)

- How changes over time are implemented
- Reporting
- Consequences/Assistance/Rewards
- Number of stages
- Aggregation rules
- Factors included
- Treatment of missing data

Upper Bar/Lower Bar

- Different application to every model
  - A and C—identify lowest performers and have them improve, then raise bar
  - B and D—exempt high performers from consequences, create separate response for low performers

Standards

- What achievement level gets mapped to what label (e.g., is “Basic” passing or is “Proficient?”)
- Percentage of students expected to meet standard
- Student level vs. school level (e.g., students need to be Basic to pass; schools need to have at least 50 percent of their students passing to be Satisfactory)

Amount of time given to meet standards

- Impact on reliability of Model B

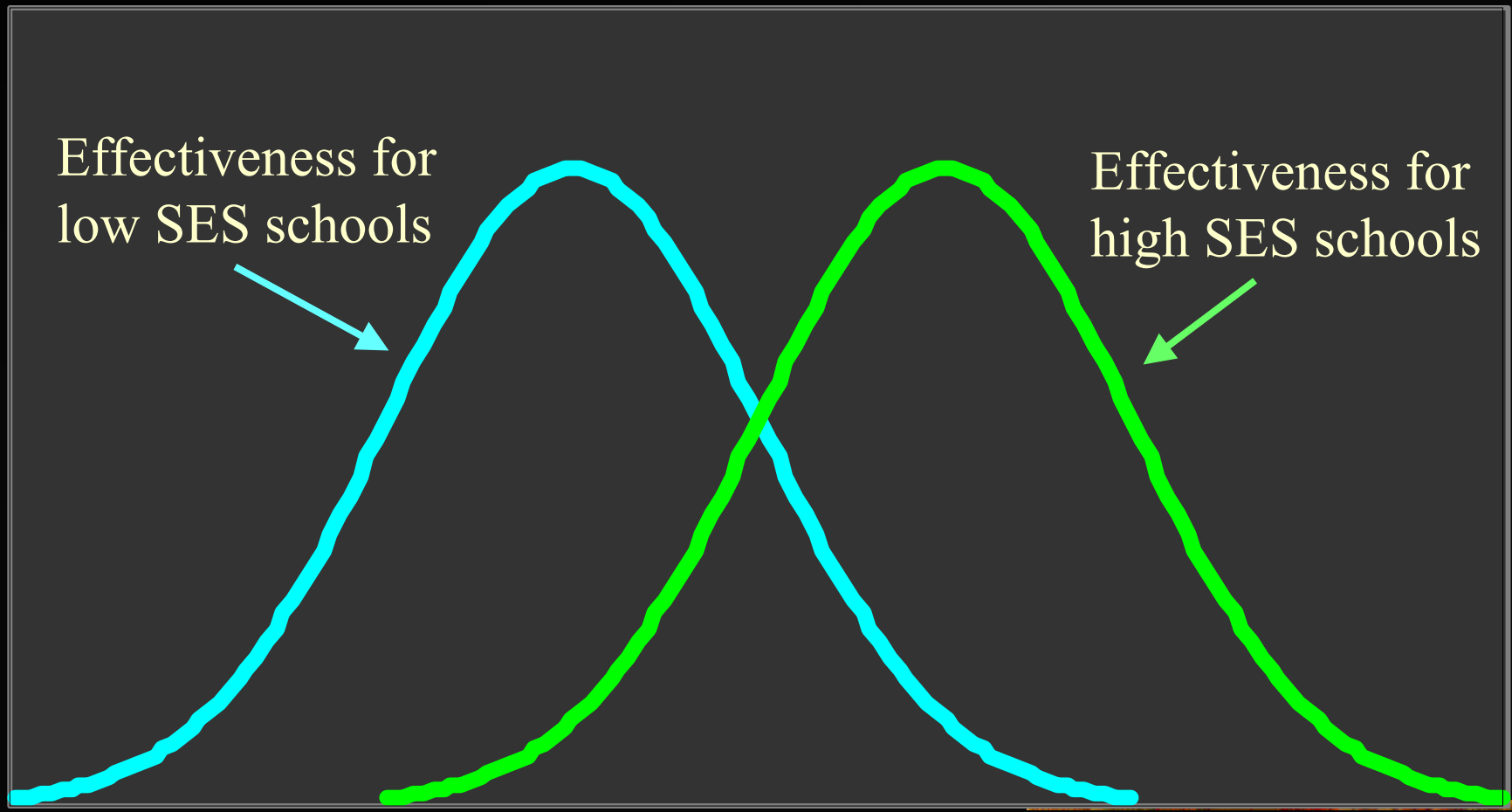
Amount of time between decisions

- Annual cycles
- Biennial cycles
- Rolling averages
- Increasing time between baseline and current status

Expectations for performance or improvement

- Current performance (3-3A, B, C, D, G, H)
- Desired performance
- Dependent on background characteristics (3-3A, B, C, D, E)

Setting Expectations for Effectiveness



Changes Over Time

- Tests/student standards
- Definition of “acceptable”
- School standards
  - Model A
  - Model C
- Students included

Reporting

- Labels
  - Schools
  - Students
- Disaggregation
- Consequences for subgroups

Consequences/Assistance/ Rewards

- How severe
  - Need to be proportionate to probability of correct classification
- How reversible
  - Money
  - Reputation
    - Schools
    - Individuals
  - Staff and student transfers
  - Student learning

Number of stages

- Decision made on accountability results
- Accountability results are an initial filter

Aggregation

- Combination rules
  - Compensatory
  - Conjunctive
- Recoding
  - Before aggregating
  - After aggregating
  - Creation of Index

Aggregation (cont'd)

- Different assessments
  - Different content areas
  - Different grades
  - Different students
- Weighting
  - Different variables
  - Different subgroups
- Years—effect of rolling average

Factors Included

- Tests
  - Grades
    - Consecutive
    - Non-consecutive
  - Content areas
  - NRT/CRT
  - Locally-determined factors

Factors Included (cont'd)

- Indicators other than tests
  - Attendance
  - Dropout
  - Others

Treatment of Missing Data

- Exempted
  - Special ed
  - LEP
- Non-exempted
- Dropouts
- Affects Model A most, by far

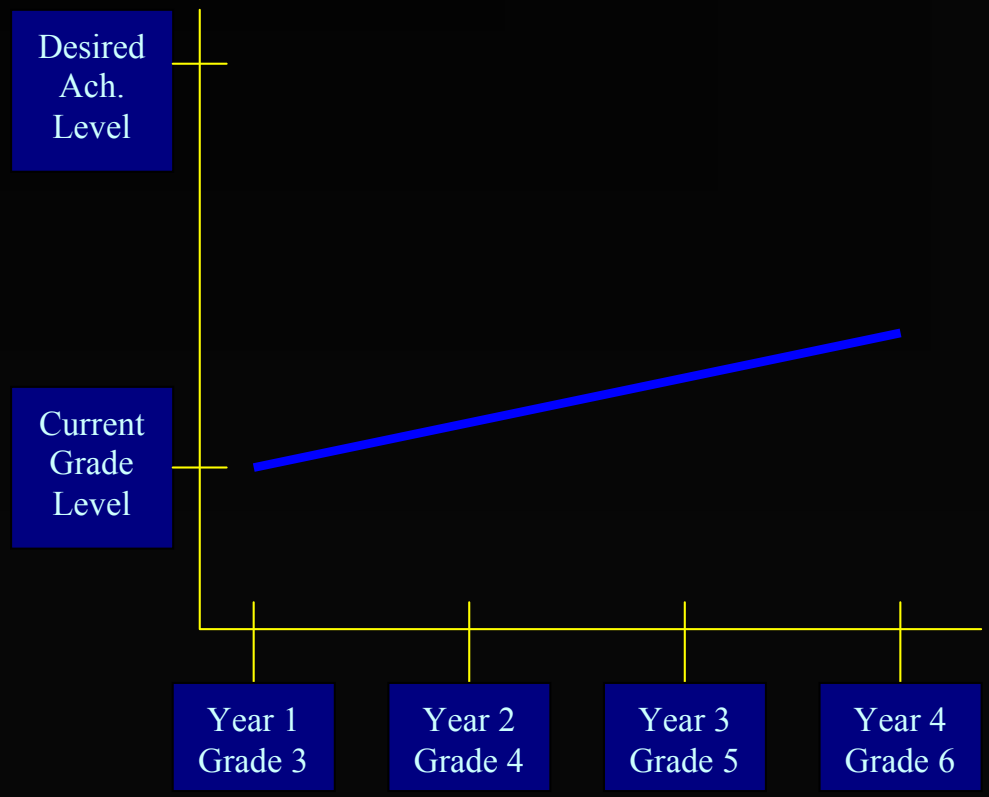
Minimum Data Requirements and Implementation Timeline

- Quad. 1—One grade, one year
- Quad. 2—One grade, two years
- Quad. 3—One cohort, two years
- Quad. 4—One cohort, three years

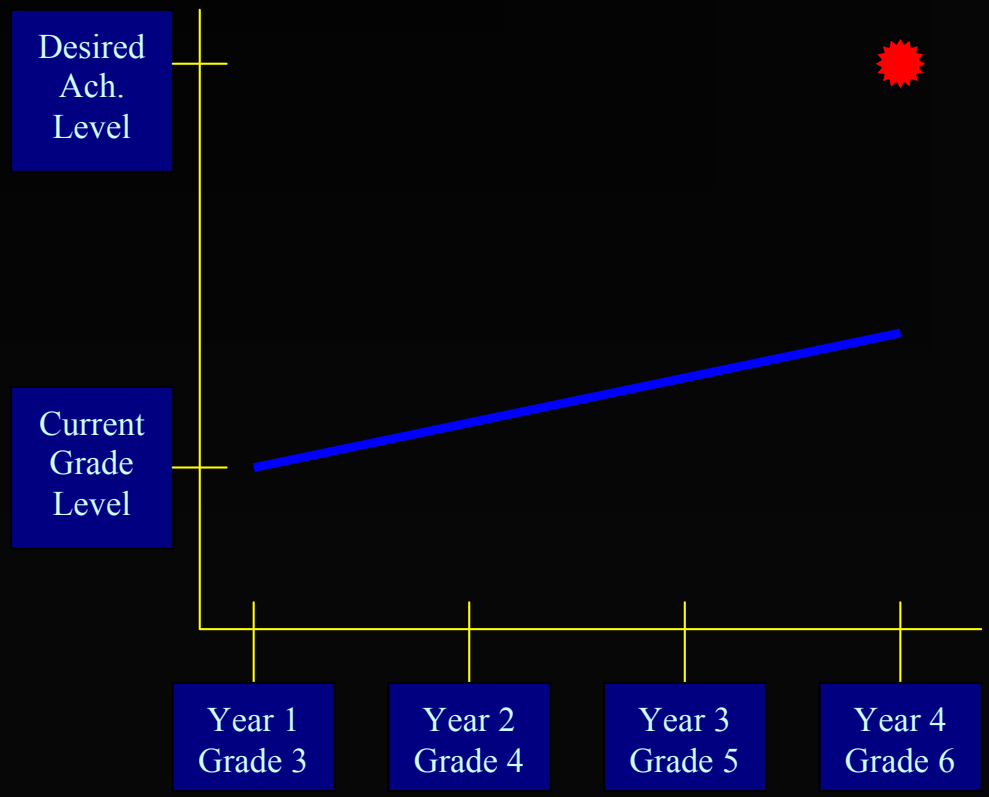
Issues/Recommendations

- Report relative to standards
- Index everything
- Plan for auditing—“Stakes changes everything”
- Dropouts
- Varying grade configurations

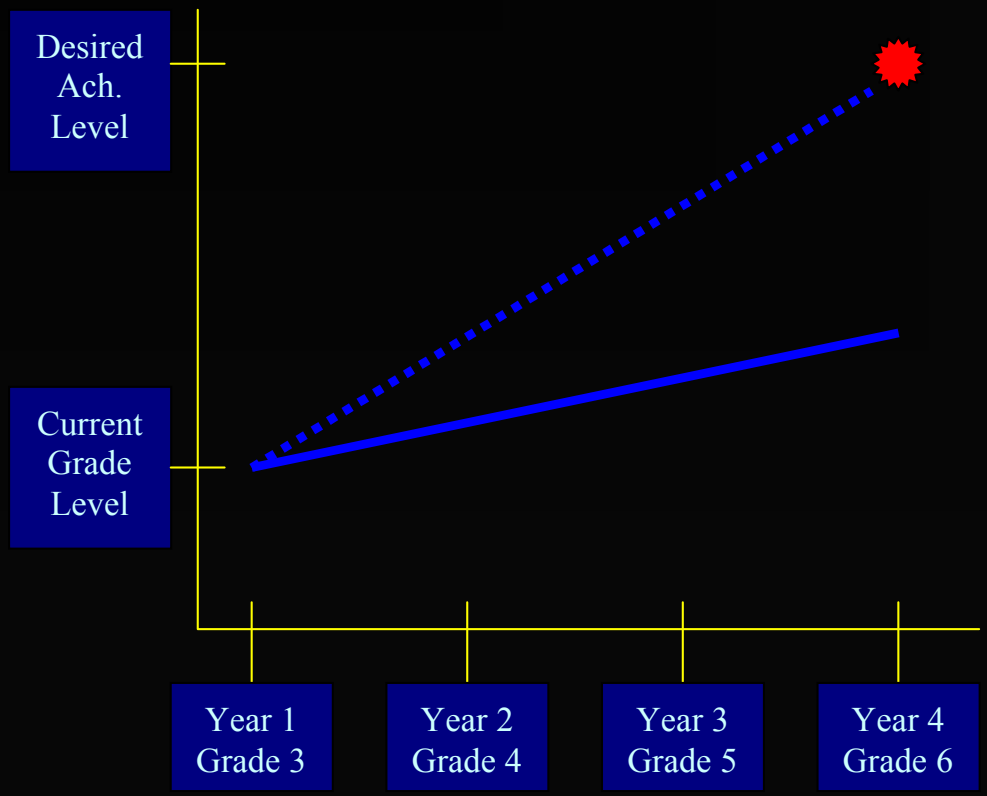
Current Situation



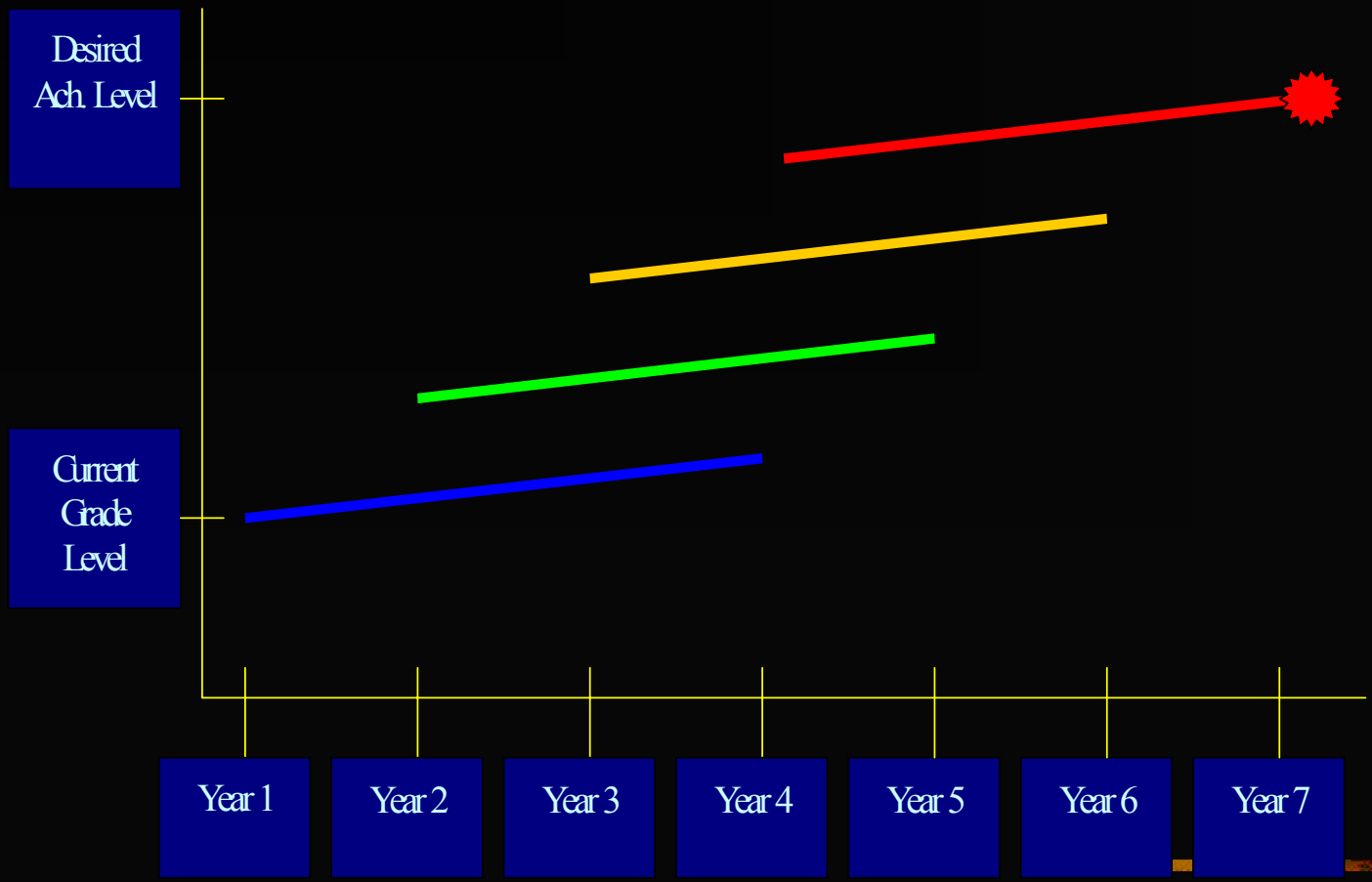
Current Situation



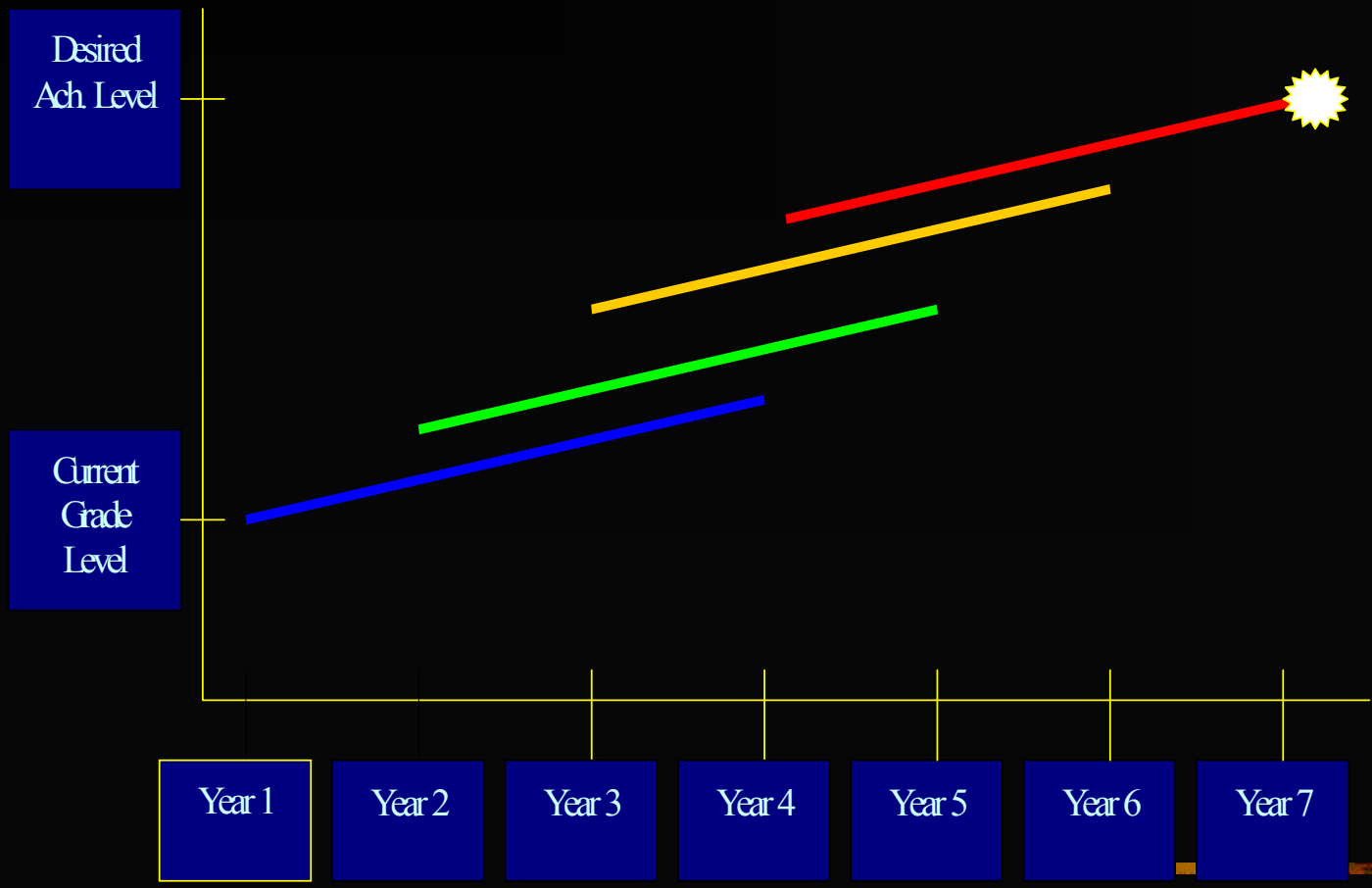
Value-Added


Desired Ach. Level
Current Grade Level
Year 1 Grade 3
Year 2 Grade 4
Year 3 Grade 5
Year 4 Grade 6

Rising Tide


Desired Ach. Level
Current Grade Level
Year 1
Year 2
Year 3
Year 4
Year 5
Year 6
Year 7

Mixed Model


Desired Ach. Level
Current Grade Level
Year 1
Year 2
Year 3
Year 4
Year 5
Year 6
Year 7

Three Choices for Improvement

- Comprehensive
  - Diffused focus
  - Limited resources
  - Small gains expected, low reliability
- Limited
  - State choice
  - Local choice