How Much School Improvement Should Accountability Systems Require?

Brian Gong and Richard Tappan

Center for Assessment

Presentation at the

2001 Reidy Interactive Lecture Series

Nashua, NH October 4-5, 2001

How much school improvement should an accountability system require?

This question is intrinsically tied to today's standards movement and school accountability systems

- Most states have established standards that require some or most students and schools to improve above where they are now
- Purpose of accountability systems is to help more students meet the state standards and have schools increase their capacities to better help students learn

Need to establish how much improvement is possible

- Some wonder whether large performance improvements can be made in public schools
- Some researchers question whether improved scores under accountability conditions are valid
- There should be some rational and empirically supported basis for setting improvement goals
- A better understanding of improvement takes place can help direct programs and policy
- Larger changes make for more reliable accountability decisions

Presentation Outline

- Define improvement or growth under four models
- Relative uncertainty of four models
- How much improvement? (state level)
- How much improvement (school level)
- Real change? Some case studies
- Thoughts about some current accountability issues
- Speculations and further questions for discussion

Presentation Outline

- Define improvement or growth under four models
- Relative uncertainty of four models
- How much improvement? (state level)
- How much improvement (school level)
- Real change? Some case studies
- Thoughts about some current accountability issues
- Speculations and further questions for discussion

Definition of Four Accountability Models

	Status	Change
Achievement	A	B
Efficiency	C	D

What is measured by each model

Grade	Year		
	1	2	3
3	A	D	G
4	B	E	H
5	C	F	I

Model	Measure in Yr 2
A	D+E+F
B	(D-A)+(E-B)+(F-C)
C	(E-A)+(F-B)
D	(In Yr 3) (H-D)-(E-A), etc.

What is measured by each model

Grade	Year		
	1	2	3
3	A	D	G
4	B	E	H
5	C	F	I

Model	Measure in Yr 2
A	D+E+F
B	(D-A)+(E-B)+(F-C)
C	(E-A)+(F-B)
D	(In Yr 3) (H-D)-(E-A), etc.

What is measured by each model

Grade	Year		
	1	2	3
3	A	D	G
4	B	E	H
5	C	F	I

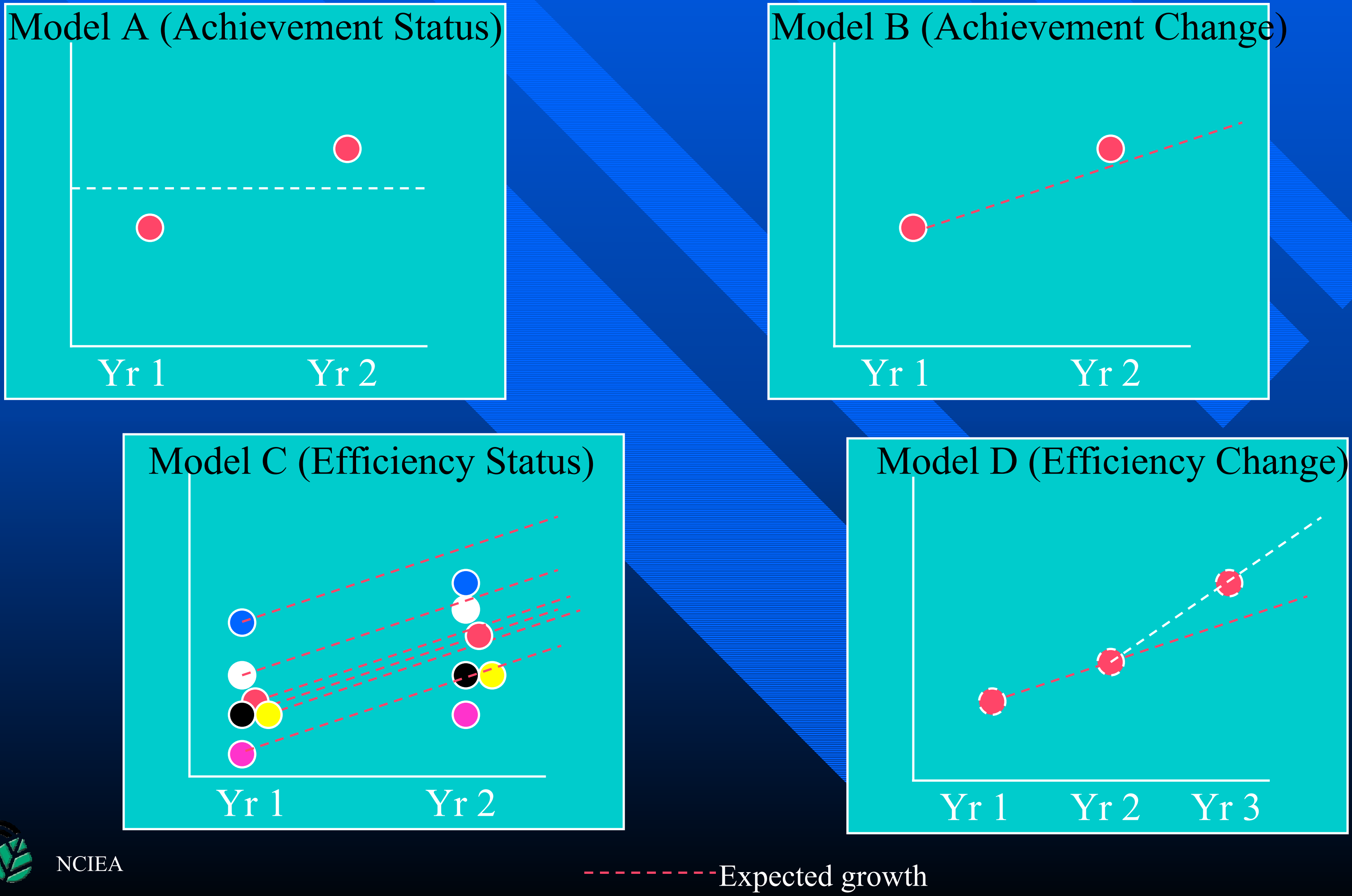
Model	Measure in Yr 2
A	D+E+F
B	(D-A)+(E-B)+(F-C)
C	(E-A)+(F-B)
D	(In Yr 3) (H-D)-(E-A), etc.

What is measured by each model

Grade	Year		
	1	2	3
3	A	D	G
4	B	E	H
5	C	F	I

Model	Measure in Yr 2
A	D+E+F
B	(D-A)+(E-B)+(F-C)
C	(E-A)+(F-B)
D	(In Yr 3) (H-D)-(E-A), etc.

Graphic view: Accountability for four models


Model A (Achievement Status)
Yr 1
Yr 2
Model B (Achievement Change)
Yr 1
Yr 2
Model C (Efficiency Status)
Yr 1
Yr 2
Model D (Efficiency Change)
Yr 1
Yr 2
Yr 3
Expected growth

What is Improvement or Growth?

Model	Measured		Observed Growth	Growth Target	Accy Decision
	Yr 1	Yr 2			
A	A,B,C	D,E,F	(D+E+F) PAC	Acceptable PAC	Did school meet acceptable PAC?
B	A,B,C	D,E,F	(D-A)+(E-B) +(F-C)	GT: Reduce Baseline to Goal/time	Meet GT; Percent Goal Achieved
C	A,B,C	D,E,F	(E-A)+(F-B)	One year's expected growth	Cutscore; Percent Goal Achieved; or # Ss
D	A,B,C	D,E,F	(H-D)-(E-A)	Accelerated improvement GT	? Percent Goal Achieved?

Examples of Four Accountability Models

	Status	Change
Achievement	A TX NC	B CA KY LA VT
Efficiency	C NC TN	D None known
Mixed A, B – OR		

Presentation Outline

- Define improvement or growth under four models
- Relative uncertainty of four models
- How much improvement? (state level)
- How much improvement (school level)
- Real change? Some case studies
- Thoughts about some current accountability issues
- Discussion and research questions

Calculating Relative Uncertainties - Approach

- To calculate school accountability decision consistency ("reliability") – see Hill, RILS 2001
- Uncertainty = amount to detect / standard error
- Ratio of uncertainties between models

Model	A	B	C	D
Amount to detect				
Standard error				
Ratio to A				

Relative Uncertainty of Four Models

- Detecting status (Model A) is 3-20 times more "certain" than change in Models B, C, or D

	A	B	C	D
Ratio of A:	1:1	7.8:1	2.8:1	20:1
	*Dependent on specific state data			

- Standard errors about the same, but
- Amount of change to be detected varies
- Each state should calculate using own data

Calculating Relative Uncertainties –Numerators

- Uncertainty = amount to detect / standard error
- Ratio of uncertainties between models

Model	A	B	C	D
Amount to detect (School SD)	1	.18	.5	.1
Standard error	1			
Ratio to A	1			

Model B Numerator

- Model B - .18 School mean SD/year
- Source: estimate from Kentucky data

Kentucky Goal: Move average student from below Apprentice (about 30th percentile in 1991) to Proficient (about 90th percentile) in 20 years

- Z-score change from -.52 to 1.29, or 1.81 student SD in 20 years, or .09 Student SD/year
- School change of .18 school SD/year; .36 for 2 years
- (our previous studies show school SD approximately ½ student SD)
- Estimates from two other states: .05, .07 school SD/yr

Model C Numerator

- Model C - .5 School mean SD/year
- Source: empirical result from a state’s data where we calculated Model C results

Model D Numerator

- Model D - .1 School mean SD/year
- Source: We estimated from Models A and B that a change of one-fifth of a school SD per year might be on the upper end of expected change. We applied the one-fifth to the .5 School SD of Model C.

Calculating Relative Uncertainties –Denominators

- Uncertainty = amount to detect / standard error
- Ratio of uncertainties between models

Model	A	B	C	D
Amount to detect (School SD)	1	.18	.5	.1
Standard error*	1	$\sqrt{2}$	$\sqrt{2}$	2
Ratio to A				

*for specified conditions

Calculating Relative Uncertainties – Notation

- Standard errors for each model – general formulas
- Standard errors for specific situation to allow comparisons

σ_X^2 = the variance of pupil observed scores within school,

$\sigma_{i\bar{X}}^2$ = the variance error of school observed mean scores for Quadrant *i*,

N = the number of students in the school accountability system each year

p = the proportion of students returning from one year to the next

r = the within-school correlation of scores from one year to the next

Calculating Relative Uncertainties – Formulas

Standard errors

Model A $$\sigma_{A\bar{X}}^{2} = \sigma_{X}^{2} / N$$

Model B $$\sigma_{B\bar{X}}^{2} = 2\sigma_{X}^{2} / N = 2\sigma_{\bar{X}}^{2}$$

Model C (TL) $$\sigma_{C\bar{X}}^{2} = 2(1 - r^{2})\sigma_{X}^{2} / pN$$

Model C (QL) $$\sigma_{C\bar{X}}^{2} = 2(1 - pr^{2})\sigma_{X}^{2} / N$$

$$\sigma_{c\bar{X}}^{2} = \{2p(1 - r^{2})\sigma_{X}^{2} + 2(1 - p)\sigma_{X}^{2}\} / N$$

Model D $$\sigma_{D\bar{X}}^{2} = 4(1 - r^{2})\sigma_{X}^{2} / pN$$

Observations about relative standard errors

Model B s.e. is twice Model A.

Looking at gain across two years doubles the error (because of sampling error associated with two groups of students)

Model B s.e. is twice-to-same-as Model C.

For TL, assuming r^2 is .5 and all student are retested, then Model C s.e. is one-half that of B. Assuming multiple grades tested (which increases the N for Model B), three-fourths of students return, and 100 students per grade for three grades, then s.e. of B and C are equal ($2\ S_x{}^2 / 300 = .0067\ S_x{}^2$).

For QL, the s.e. for the values above would be slightly lower ($5\ S_x{}^2 / 800 = .0063\ S_x{}^2$).

Standard Errors for Example

Example: K-5 schools, with 3 grades tested (3, 4, 5), equal numbers of students per grade, two-thirds of students return from previous year, and r^2 of scores between years is .5

Example Standard Errors for Four Models

Standard errors

Model A $\sigma_X^2 / N \rightarrow \sigma_X^2 / 3N$

Model B $2\sigma_X^2 / N \rightarrow 2\sigma_X^2 / 3N$

Model C$_{\text{(TL)}}$ $2(1-r^2)\sigma_X^2 / pN \rightarrow 2(.5)\sigma_X^2 / \tfrac{2}{3}\,2N$

Model D $4(1-r^2)\sigma_X^2 / pN \rightarrow 2(.5)\sigma_X^2 / \tfrac{2}{3}\,2N$

Example Standard Errors for Four Models - continued

Standard errors

Model A $\sigma_X^2 / 3N$

Model B $2\sigma_X^2 / 3N \rightarrow 2\text{variance} \rightarrow \sqrt{2}SD$

Model C$_{\text{(TL)}}$ $2(.5)\sigma_X^2 / \tfrac{2}{3}\, 2N \rightarrow \approx \sqrt{2}SD$

Model D $4(.5)\sigma_X^2 / \tfrac{2}{3}\, 2N \rightarrow \approx 2SD$

Calculating Relative Uncertainties – Ratios

- Uncertainty = amount to detect / standard error
- Ratio of uncertainties between models

Model	A	B	C	D
Amount to detect* (School SD)	1	.18	.5	.1
Standard error*	1	$\sqrt{2}$	$\sqrt{2}$	2
Ratio to A*	1	7.9	2.8	20

*for specified conditions

Relative Uncertainties – Observations

Model	A	B		C	D
		1 yr	2 yr		
Amount to detect* (School SD)	1	.18	.36	.5	.1
Standard error*			$\sqrt{2}$	$\sqrt{2}$	2
Ratio to A*	1	7.9	3.8	2.8	20

*for specified conditions

- To equal level of uncertainty of Model A:
  - Model B would need to increase (2 year) amount to detect 3.8 times, or decrease standard error 3.8 times
  - Model C would need to increase/decrease 2.8 times

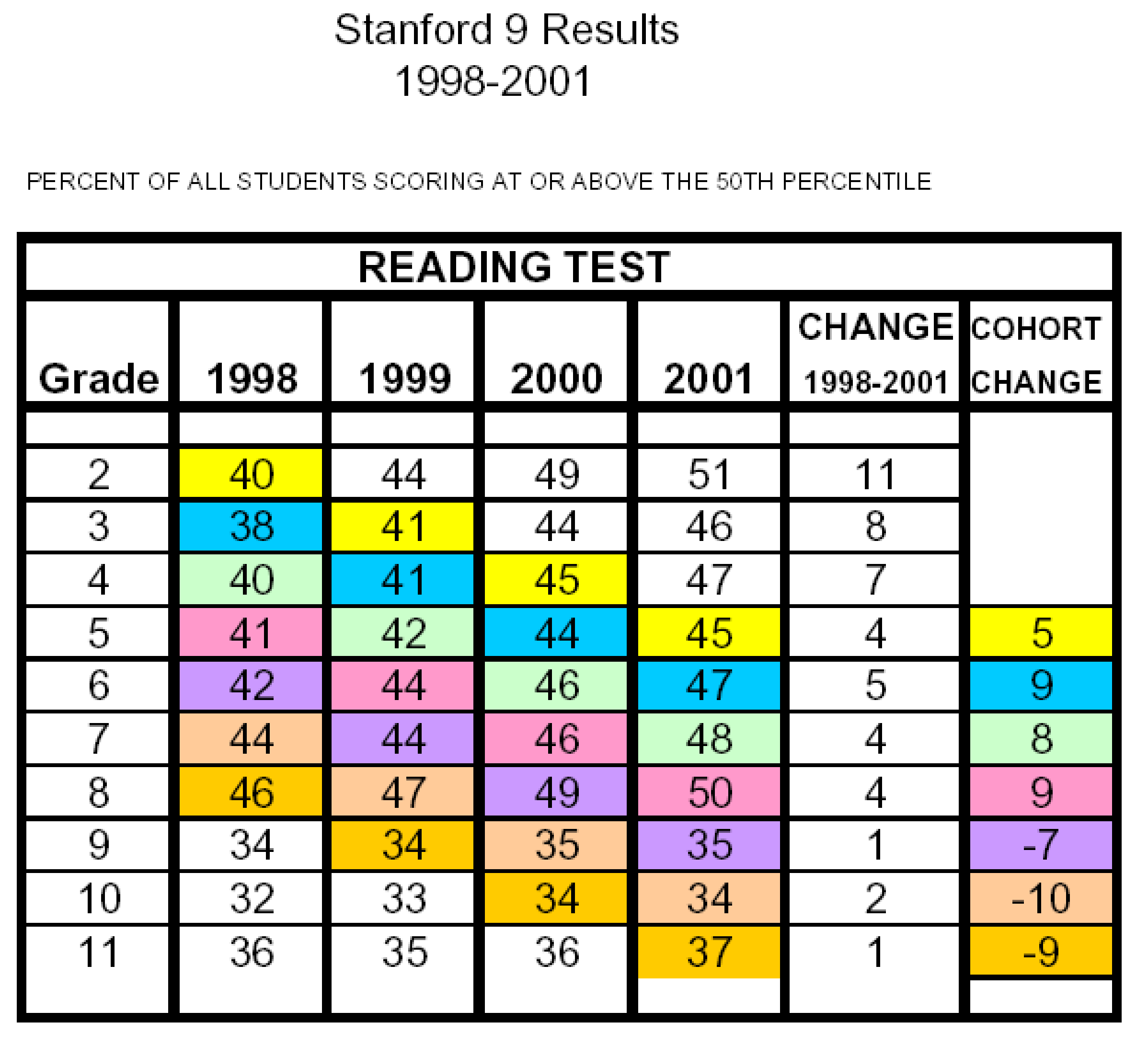
To decrease uncertainty

- Is the uncertainty/inconsistency/ unreliability appropriate and acceptable?
- Can decrease standard error
  - Decrease student sampling error
    - » Increase numbers of different students tested
- Can increase amount to be detected

Presentation Outline

- Define improvement or growth under four models
- Look at relative reliability of four models
- How much improvement? (state level)
- How much improvement (school level)
- Real change? Some case studies
- Thoughts about some current accountability issues
- Speculations and further questions for discussion

Example: California



Stanford 9 Results
1998-2001

PERCENT OF ALL STUDENTS SCORING AT OR ABOVE THE 50TH PERCENTILE

READING TEST

Grade	1998	1999	2000	2001	CHANGE 1998-2001	COHORT CHANGE
2	40	44	49	51	11	
3	38	41	44	46	8	
4	40	41	45	47	7	
5	41	42	44	45	4	5
6	42	44	46	47	5	9
7	44	44	46	48	4	8
8	46	47	49	50	4	9
9	34	34	35	35	1	-7
10	32	33	34	34	2	-10
11	36	35	36	37	1	-9

Example: North Carolina - PAC

Reading Test

percent of all students scoring at or above minimum passing score (Levels III, IV)

Grade	1994	1995	1996	1997	1998	1999	2000	CHANGE 1994-2000		COHORT CHANGE	
								94-97	97-00	94-97	97-00
3	60	63	65	66	72	74	74	6	8		
4	66	64	69	68	71	71	72	2	4		
5	66	68	67	71	75	76	79	5	8		
6	65	66	68	67	70	72	70	2	3	7	4
7	64	69	67	68	71	77	76	4	8	2	8
8	71	73	73	75	80	80	83	4	8	9	12

Example: Texas – PAC

Reading Test

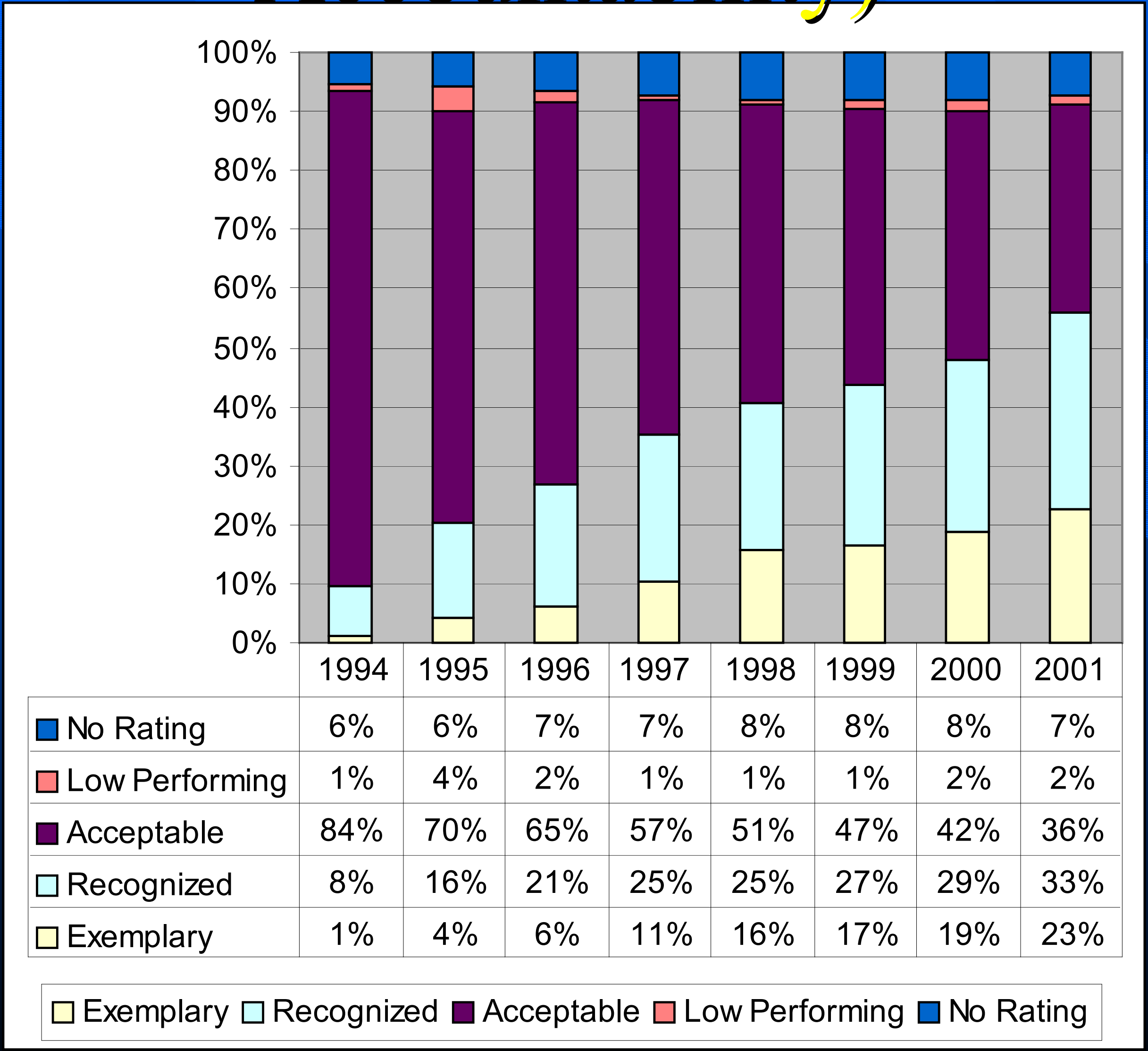
percent of all students scoring at or above minimum passing score

Grade	1994	1995	1996	1997	1998	1999	2000	CHANGE 1994-2000		COHORT CHANGE	
								94-97	97-00	94-97	97-00
3	76	77	78	78	83	88	87	2	9		
4	73	78	75	79	86	88	89	6	10		
5	75	77	79	81	85	86	87	6	6		
6	71	76	74	81	82	84	86	10	5	5	8
7	73	76	79	81	82	83	83	8	2	8	4
8	74	76	79	81	82	83	83	7	2	6	2
10	74	72	74	80	81	88	89	6	9	9	8

How much improvement should be expected?

- "Ought" goals usually not "what is"
  - Need link to a standards-based end-point
  - Variable from year-to-year (using PAC; Model B, C, D)
  - Changes under accountability conditions difficult to interpret
- May see large changes over time
  - Valid changes?
  - State average may not be appropriate estimate for change at school level
  - School changes may not happen within same small windows

Example: Texas (Model A School Accountability)



	1994	1995	1996	1997	1998	1999	2000	2001
No Rating	6%	6%	7%	7%	8%	8%	8%	7%
Low Performing	1%	4%	2%	1%	1%	1%	2%	2%
Acceptable	84%	70%	65%	57%	51%	47%	42%	36%
Recognized	8%	16%	21%	25%	25%	27%	29%	33%
Exemplary	1%	4%	6%	11%	16%	17%	19%	23%

50% Ss

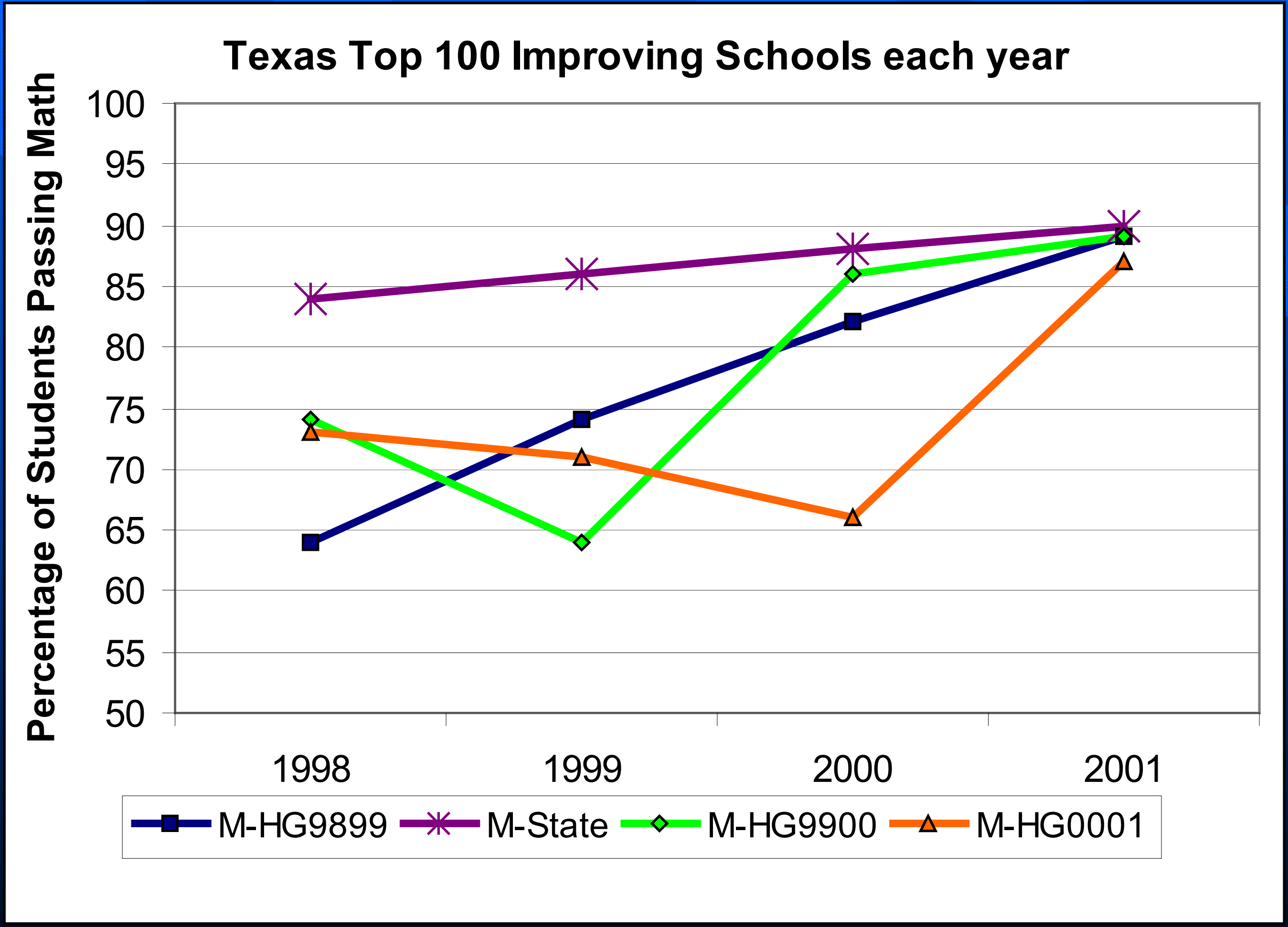
80% Ss

90% Ss

Presentation Outline

- Define improvement or growth under four models
- Look at relative reliability of four models
- How much improvement? (state level)
- How much improvement (school level)
- Real change? Some case studies
- Thoughts about some current accountability issues
- Speculations and further questions for discussion

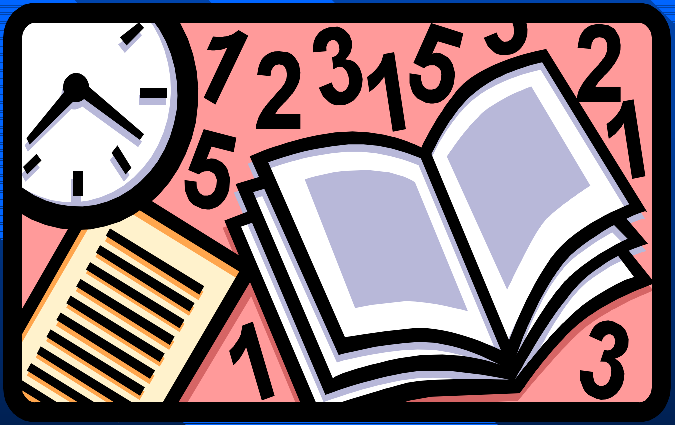
Some schools "turn on" substantial score gains at different times



Presentation Outline

- Define improvement or growth under four models
- Look at relative reliability of four models
- How much improvement? (state level)
- How much improvement (school level)
- Real change? Some case studies
- Thoughts about some current accountability issues
- Speculations and further questions for discussion

School Improvement: Cases in Major Change



Richard Tappan
Center for Assessment
Portsmouth, NH
October 2001

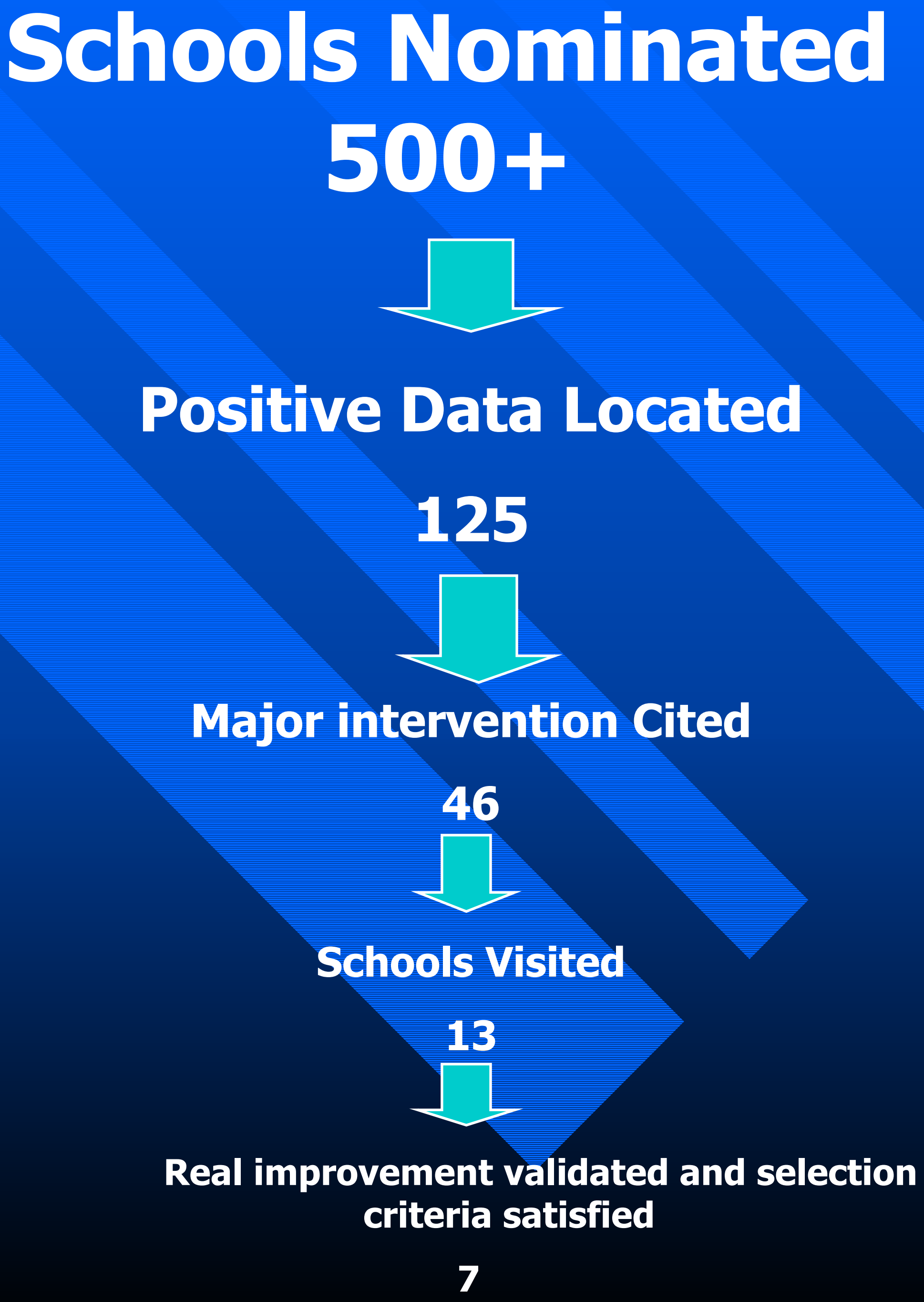
The Purpose of the Improved Schools Project

To answer these questions:

- Are the apparent gains the function of *real* change in learning?

- How much can an effective school improve in *real learning* from one year to the next?

- What are common characteristics of highly improved schools?



Schools Nominated 500+

Positive Data Located

125

Major intervention Cited

46

Schools Visited

13

Real improvement validated and selection criteria satisfied

7

What was the protocol of our visits?

1. Brief meeting with administrator(s) to discuss:

   --details of the visit (whose classrooms, why those classes were selected)

   --clarification about some data provided or questions posed prior to the visit

2. Classroom visits (partial or full period depending on the tasks under way)

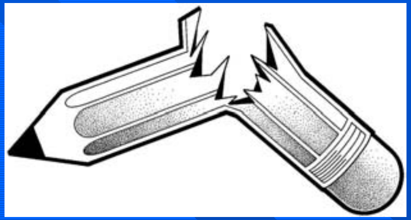
3. Visits with groups of teachers (during prep periods or after school)

4. A general tour of the building with opportunities to interview some faculty and students randomly

Criteria for our Selection

- ❖“Exemplary growth” sustained over multiple years in reading and/or math
- ❖Previously low-performing over several years
- ❖High percentage on free/reduced lunch (35%+) usually paralleled by high percentage of minorities
- ❖Major intervention cited, verified
- ❖Major change evidenced
- ❖Sound instructional practices observed

Schools visited but not cited



A Maine high school

- Growth not sustained
- Systemic changes deteriorated
- Inconsistent quality in instructional practices

3 New York City Public Schools

- Selectivity based on motivation

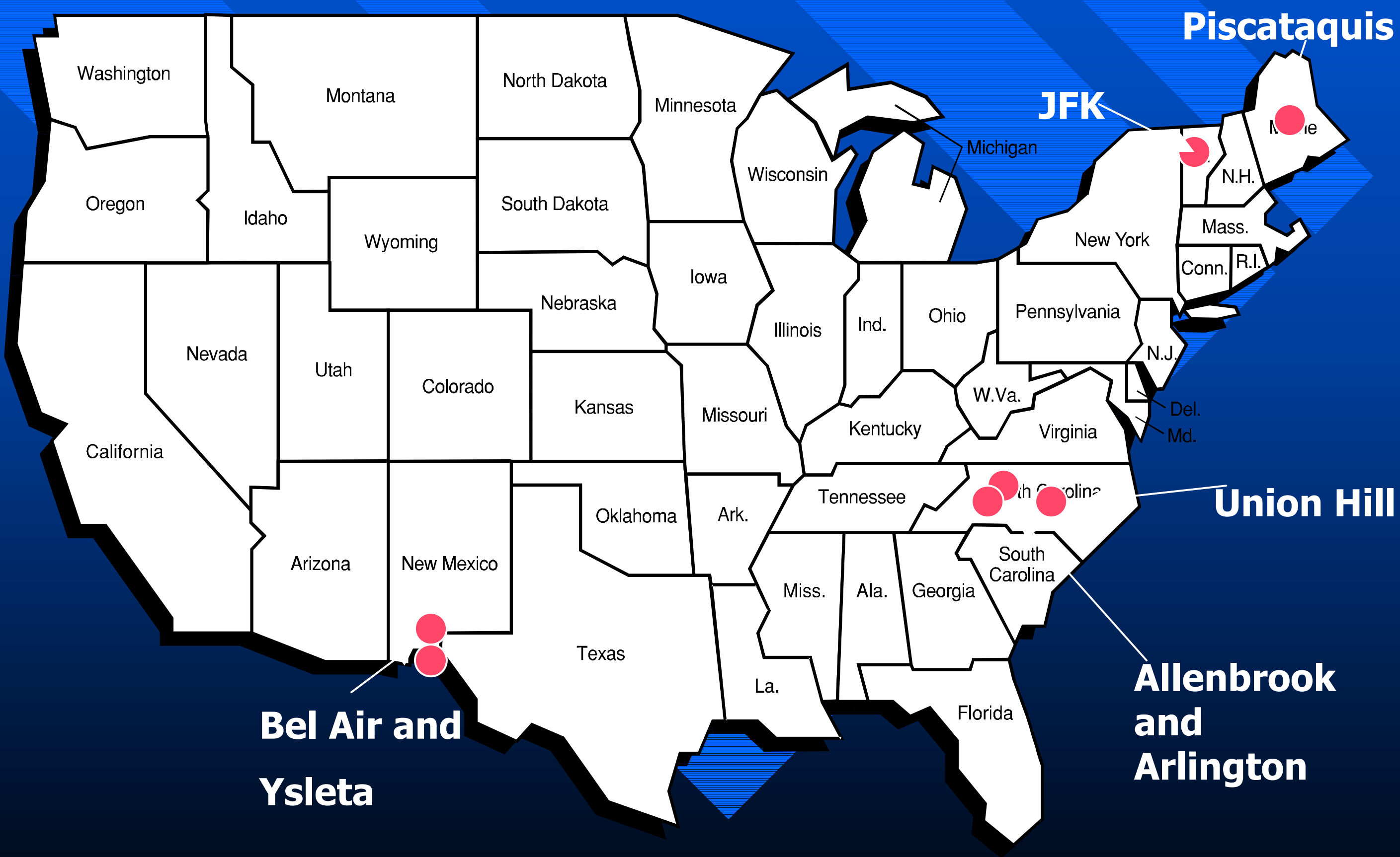
An El Paso elementary school

- Test preparation was the focus of curriculum and instruction

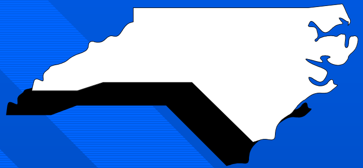
A North Carolina elementary school

- Inadequate data to confirm change

The Seven Schools Cited

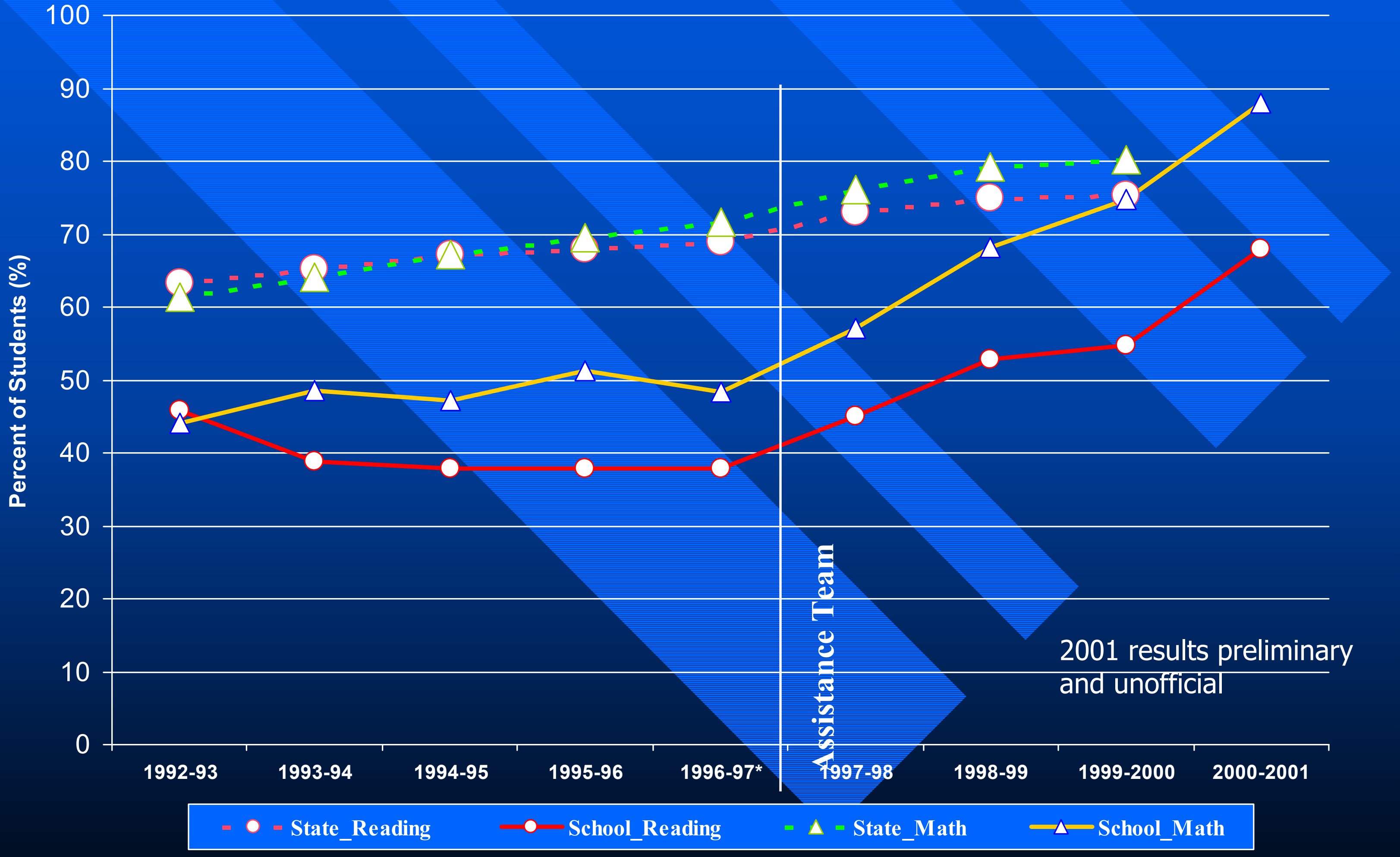


Allenbrook Elementary School Charlotte, NC



Enrollment: 326 students grades K-5; 83% minority, 17% white with African-Americans representing 65% of the total enrollment. 69% of the students are on free/reduced lunch.
Community context: Extreme transience. Only 17% of 3rd graders were there in kindergarten.
Special program: Extensive faculty mentoring. No faculty turnover for 2001-2 school year.
Major Intervention spring 1997: State declared Allenbrook "low performing." Assistance team assigned, new principal appointed; major systemic overhaul.
Status in 2000: Exemplary growth.

Allenbrook before and after intervention



2001 results preliminary and unofficial

Bel Air High School El Paso, TX

Enrollment: 2,154.

Demographics: 95% Hispanic; **83%** Free/reduced lunch; 1% Drop out rate.

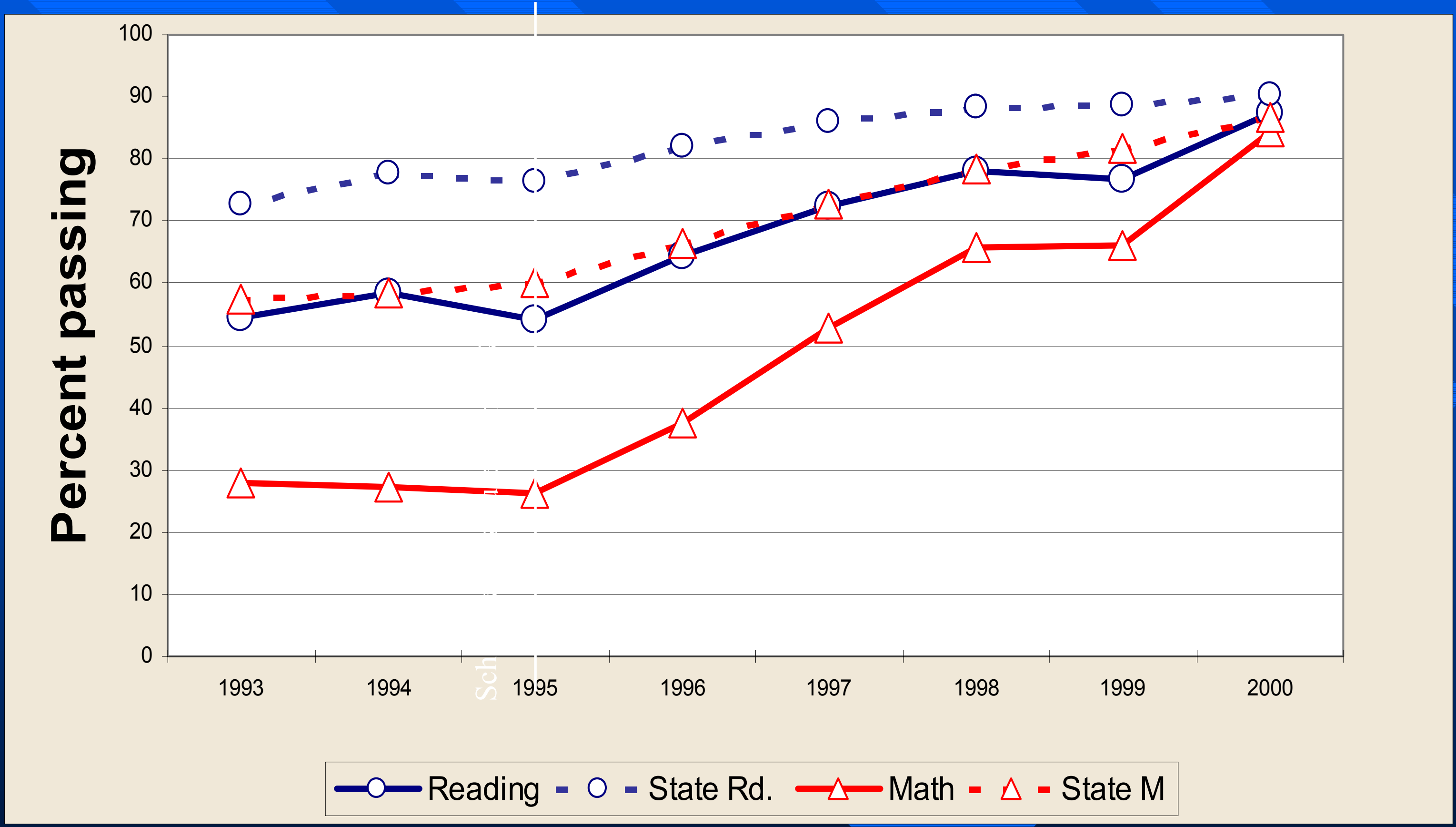
Special Recognition: 2001 “Inspiration Award” from Educational Testing Service for a 135% increase in number of students taking AP tests. 481 students took AP tests in 2000. Blue Ribbon School for 2000.

Special Program: Enrichment and remedial summer school: 50% attend; business partnership in developing curriculum.

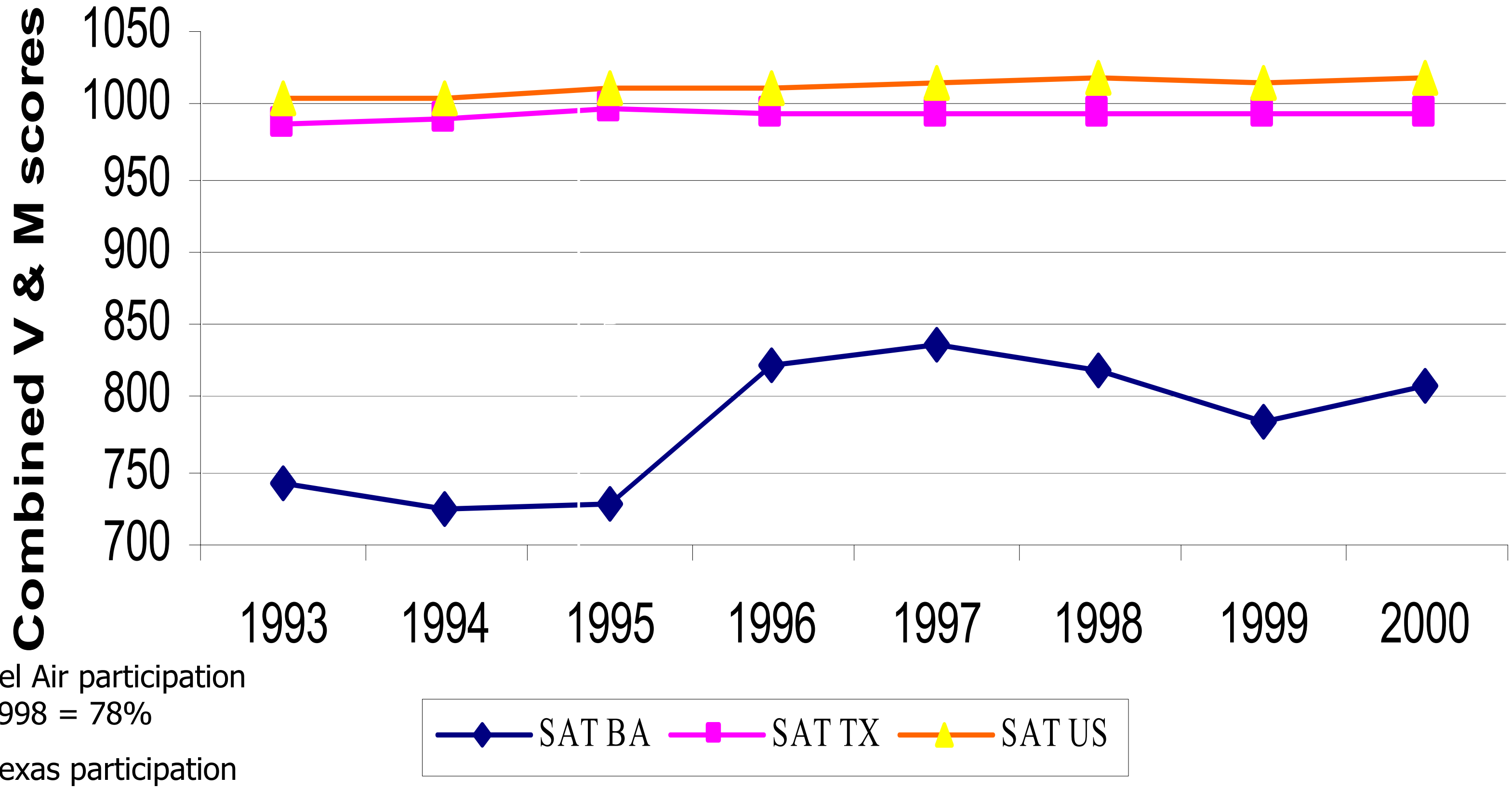
Major Intervention beginning January 1996: School reconstituted. Changes in professional development, curriculum, schedule, parent and community involvement, ongoing assessment based on school wide data.

Status in 2000: “Recognized.” Over 80% must pass the Texas assessment in all tests and in all student groups to achieve “recognized” status.

Bel Air TAAS Scores v. State



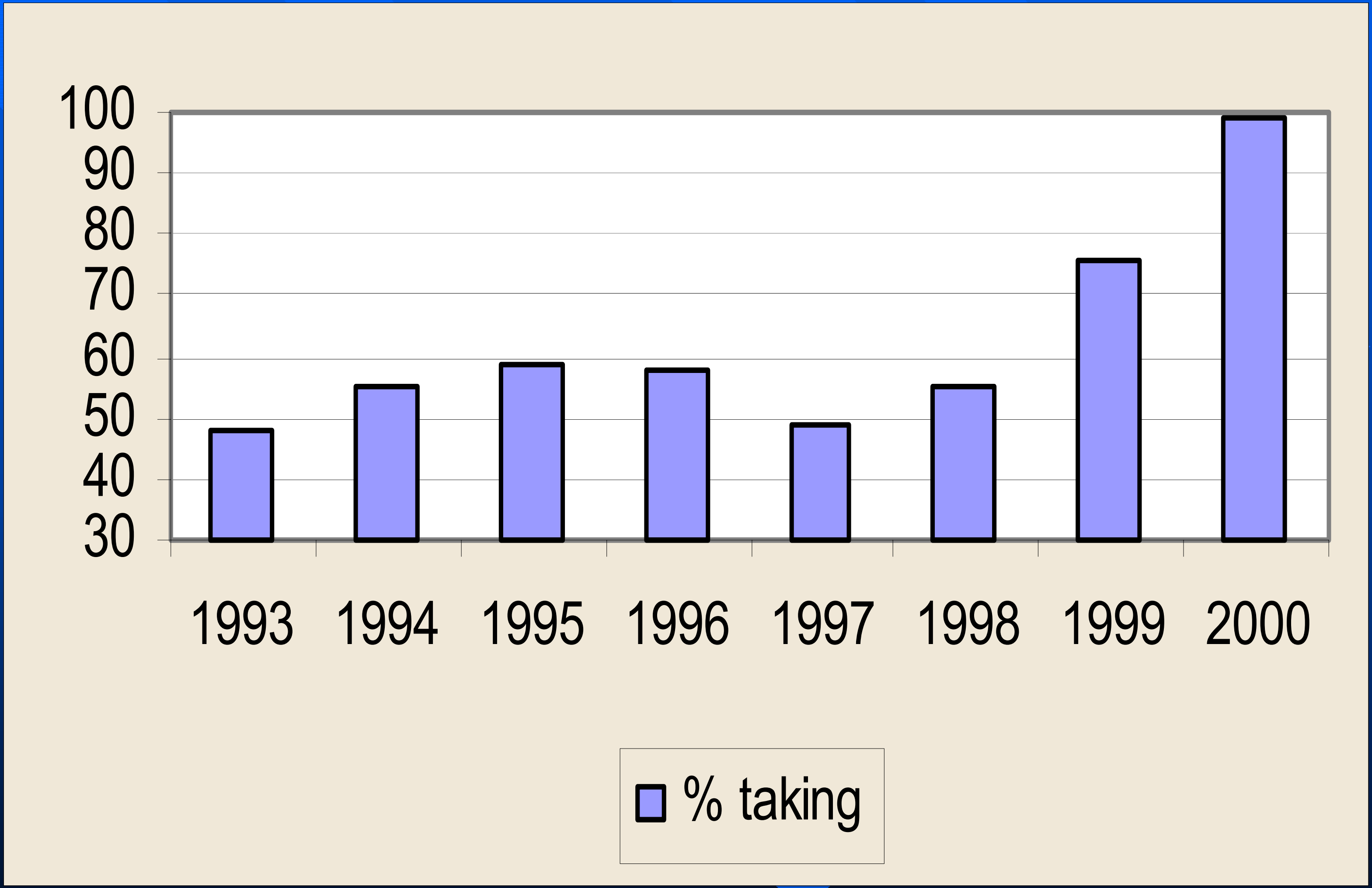
Bel Air SAT Scores v. Texas and US



Bel Air participation
1998 = 78%

Texas participation
1998 = 63%

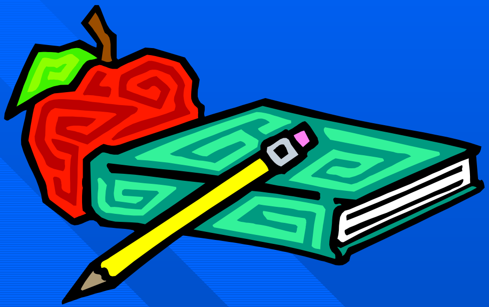
Bel Air Percentage Taking SAT



NCIEA

Common Characteristics

Numbers represent schools listed; black=characteristic observed; green=characteristic not observed



1. Strong principal with vision, will, support

1 2 3 4 5 6 7

2. New principal who comes in as agent of major change

1 2 3 4 5 6 7

3. Consistent expectations of all students from one classroom to the next

1 2 3 4 5 6 7

4. Major outside influence helped initiate change

1 2 3 4 5 6 7

5. Establishment of uniform instructional policies

1 2 [✓ 3] 4 5 [✓ 6] 7

6. Maximized time on task

1 2 3 4 [✓ 5] [✓ 6] 7

7. Resources directed at focused objectives

1 2 3 4 5 6 7

8. Frequent focused observations of teachers reported

1 2 3 4 5 6 7

9. Replacement of significant percentage of teachers within 1 or 2 years of major reform

1 2 3 4 5 6 7

10. Regular use of data on student performance to adjust instruction and assess effectiveness of the program

1 2 3 4 5 6 7

11. An inclusive “can do” atmosphere promoting high expectations as a school community

1 2 3 4 5 6 7

12. Comprehensive campaign to communicate new vision to all stakeholders

1 2 3 4 5 6 7

13. New facility or major renovations at time of major intervention strategy

1 2 3 4 5 6 7

14. Major effort to tidy, beautify and organize physical surroundings

1 2 3 4 5 6 7

15. High correlation between articulated vision
and daily practice through ongoing classroom assessment

1 2 3 4 5 6 7

16. Highest priority placed on reading and math skills

1 2 3 4 5 6 7

17. Curriculum aligned with state standards

1 2 3 4 5 6 7

18. Comprehensive, aggressive program of parental involvement in educational process

1 2 3 4 5 6 7

19. Effective involvement of community organizations

1 2 3 4 5 6 7

20. Professional development integrated with school-wide objectives, vision and individual teacher performance

1 2 3 4 5 6 7

Characteristics of all seven of the schools we cited

- A strong, visionary principal with support

- A major outside impetus for change
- Resources focused on school-wide academic objectives
- Regular use of data to adjust instruction
- Frequent focused observation of teachers

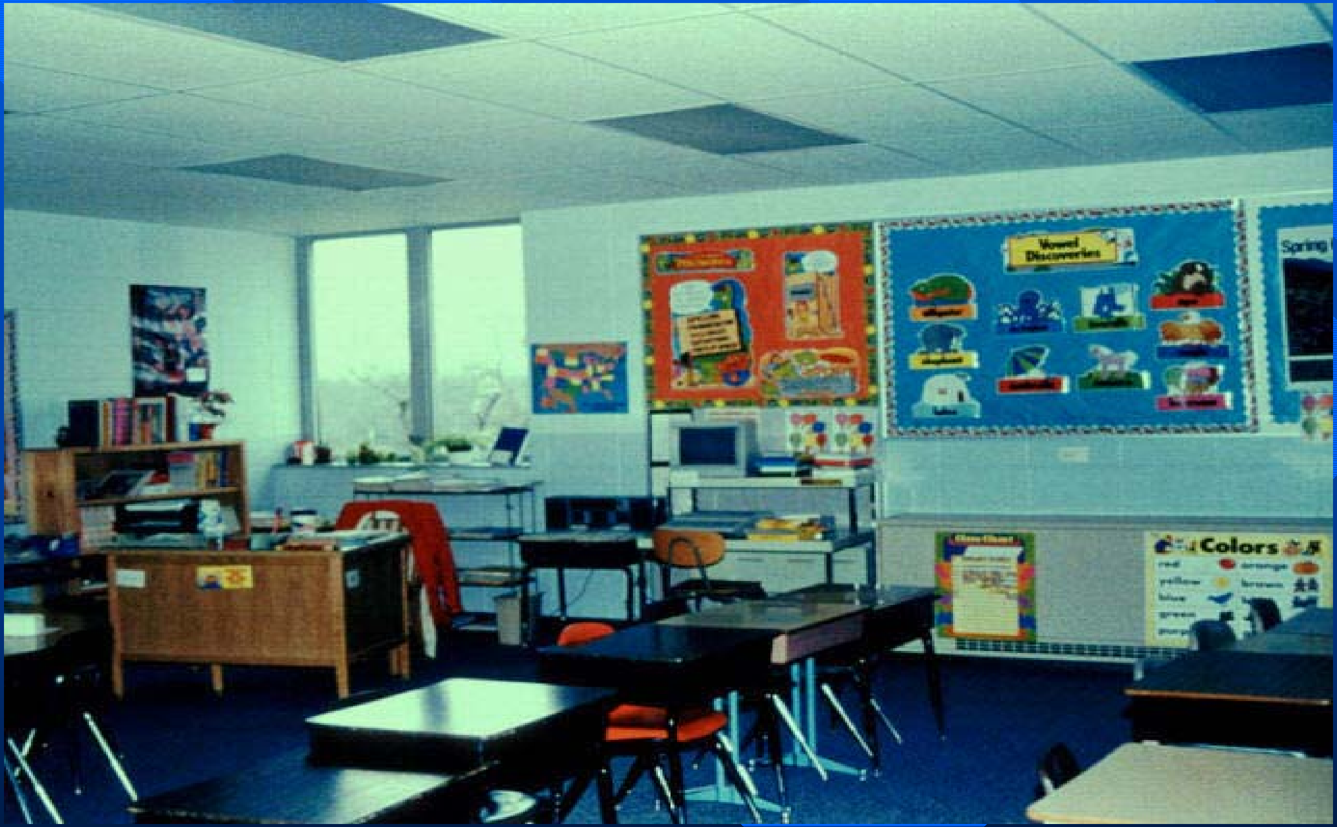
- Inclusive “can do” atmosphere
- High correlation between vision and practice
- Community and/or parental involvement
- Professional development related to school-wide objectives

What have we learned thus far from 7 schools?

Major change. . .

1. is possible in as few as 2-3 years, although difficult and rare.
2. can be sustained for years after the initial intervention.
3. can survive the departure of a dynamic principal if that person established ownership by stakeholders in the new vision.
4. is generally heralded by visible improvements in the physical surroundings in terms or order, maintenance, and aesthetics.
5. is accompanied by consistency in instruction, high level of communication, frequent assessment, targeted resources, and adjustment of program.
6. is evidenced in systemic culture change recognized by all stakeholders who can articulate past changes and current focus.

The source of the best evidence of major change is the chief product of the classroom:



Student work.

Presentation Outline

- Define improvement or growth under four models
- Look at relative reliability of four models
- How much improvement? (state level)
- How much improvement (school level)
- Real change? Some case studies
- Thoughts about some current accountability issues
- Speculations and further questions for discussion

Adequate (Yearly) Progress

- Link desired progress to longer-term goals
- Trying to detect *yearly* progress will be limited by low reliability
  - Approaches should be examined such as looking at multiple years, or Kentucky's current model which makes decisions more reliable as time goes on
- Consider using a "normalized required progress" to look across states' standards and demand for improvement
- More research on school rates of improvement different from state averages
- Attend to validity in design and implementation

“Narrowing the Gap”

- Distinguish between
  - – Having all students/schools meet a common standard
    - » Can happen at different times and rates
  - – Having all students/schools achieve the same absolute scale score
    - » Requires convergence of performance and time; requires different rates, efficiencies
      - Rates for schools can be linear (e.g., Kentucky)
      - Implications of student-level convergence?

Presentation Outline

- Define improvement or growth under four models
- Look at relative reliability of four models
- How much improvement? (state level)
- How much improvement (school level)
- Real change? Some case studies
- Thoughts about some current accountability issues
- Discussion and research questions

Discussion and research questions

- What "accountability models" will we, educators, policymakers, or the public value? How can we further a thoughtful dialogue?
  - How much uncertainty is acceptable?
  - How much "un-validity" is acceptable? What validity-reliability trade-offs are appropriate? Inappropriate? Other accountability approaches to resolve these problems?
- How widespread are large improvements in school performance? (measured by Models A, B, C)
  - How can states, districts, schools, communities, and partners "scale up" to support large, sustained improvements in student performance and school/district capacity?

Contact Information

Brian Gong and Dick Tappan

bgong@nciea.org rtappan@nciea.org

The Center for Assessment
www@nciea.org
P.O.Box 4084
Portsmouth, NH 03820
(603) 766-7900