School Classification Error

Richard Hill and Charles DePascale
Center for Assessment
June 25, 2003
LSA Annual Conference

Elements of NCLB Designs

- Outcome criterion is percent passing
- School as a whole and every subgroup within the school must pass either a status bar or an improvement standard on reading and math to make AYP
- A school that fails to make AYP two consecutive years faces serious consequences

Reliability

- Probability of a consistent or correct decision (not a reliability coefficient)
- One negative error for any subgroup within a school on either test misclassifies the whole school
- Inference is to larger population
- Results for a school or subgroup can vary considerably from year to year—similar to random draws from school's population

Point to Note

- Sampling error, not measurement error, is primary factor
  - Example: N = 50, SD = 100, r = .80
    - SE with measurement error only = 6.3
    - SE with sampling error only = 14.1
    - SE with sampling and measurement error = 15.9

Reliability of School Means

N	Test Reliability	School Mean Reliability
25	.60	.82
	.90	.89
50	.60	.90
	.90	.94
100	.60	.95
	.90	.97

Reliability Studies

- 4 Methods
  - Direct Computation
  - Split-Half
  - Monte Carlo
  - Sampling with Replacement ("bootstrapping")
- For details, see "Determining the Reliability of School Scores"

Quick Study to Demonstrate Accuracy of Assumptions

- Assumption of random draws of students allows us to calculate, for example, standard deviation of difference scores
- For example, standard deviation of difference scores when N = 50 is predicted to be 10, when N = 100, 7.1
- How much actual variation is there compared to what the equations predict?

Quick Study to Demonstrate Accuracy of Assumptions

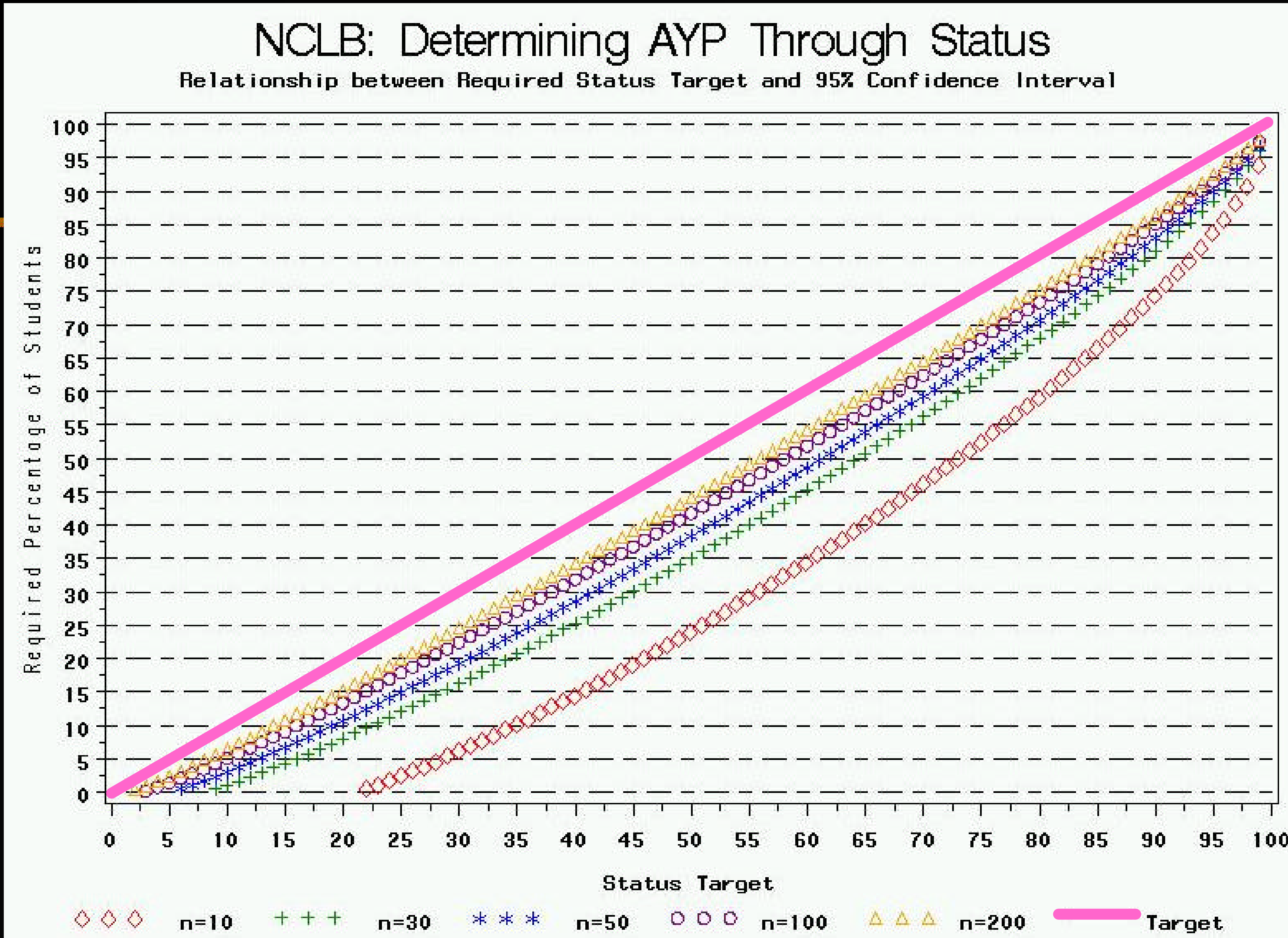
- Could compute the standard deviation of differences in schools' percent proficient across years, but that would be confounded with changes in the educational programs
- Computed the difference between the percentage of males in 2001 and 2002

Comparison of Predicted SD to Actual SD

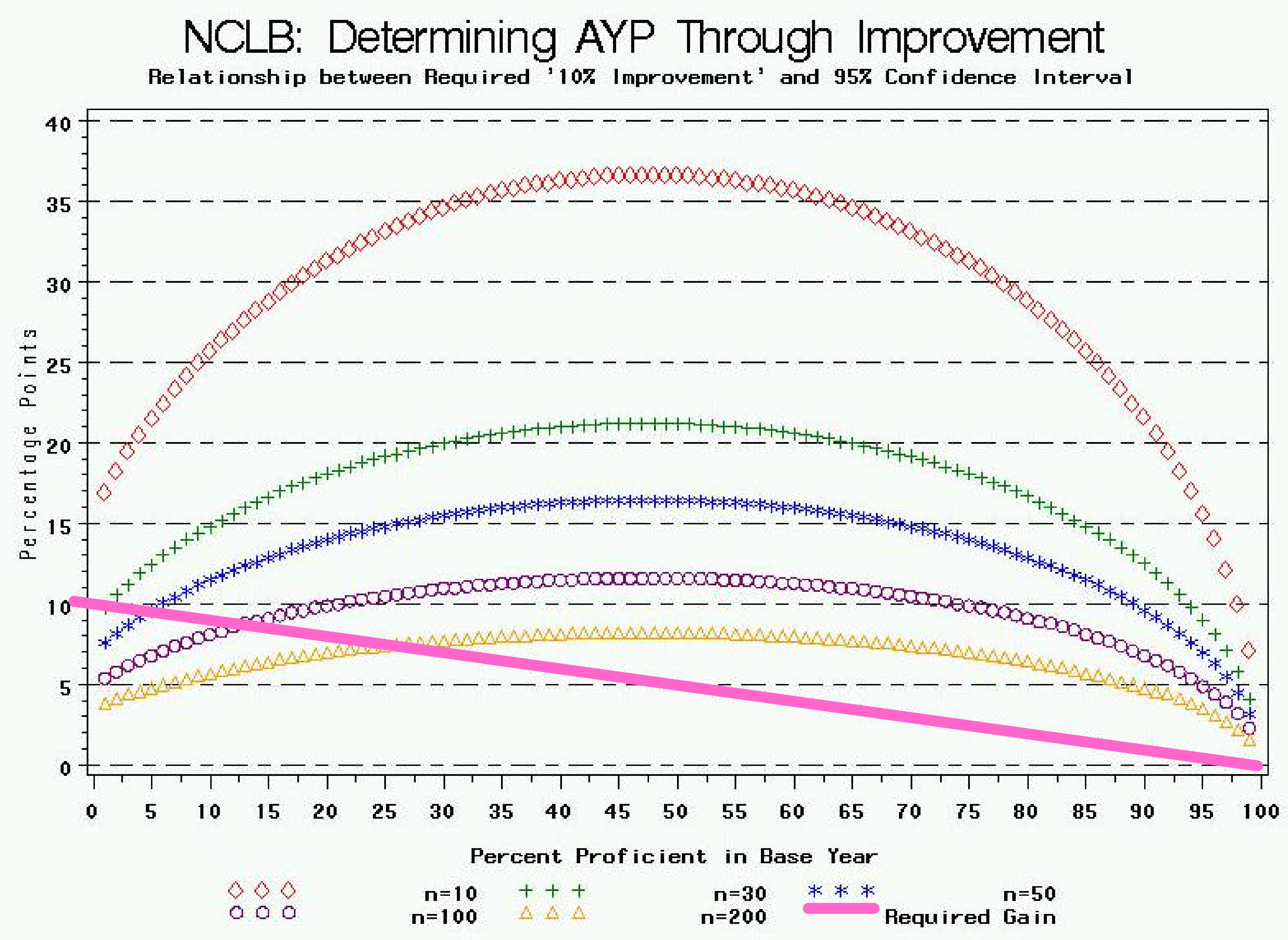
	Pre-dicted	Actual	Pre-dicted	Actual
Number of Students/School	50	40-60	100	80-120
Standard Deviation of Differences	10	10.3	7.1	7.3

Status vs. Improvement

- Generally can relatively reliably determine status with groups of moderate size
  - One year of error
  - Subgroups often are far from 20^{th} %tile school
- Generally cannot reliably determine improvement even with very large groups
  - Two years of error
  - Amount of improvement expected is relatively small



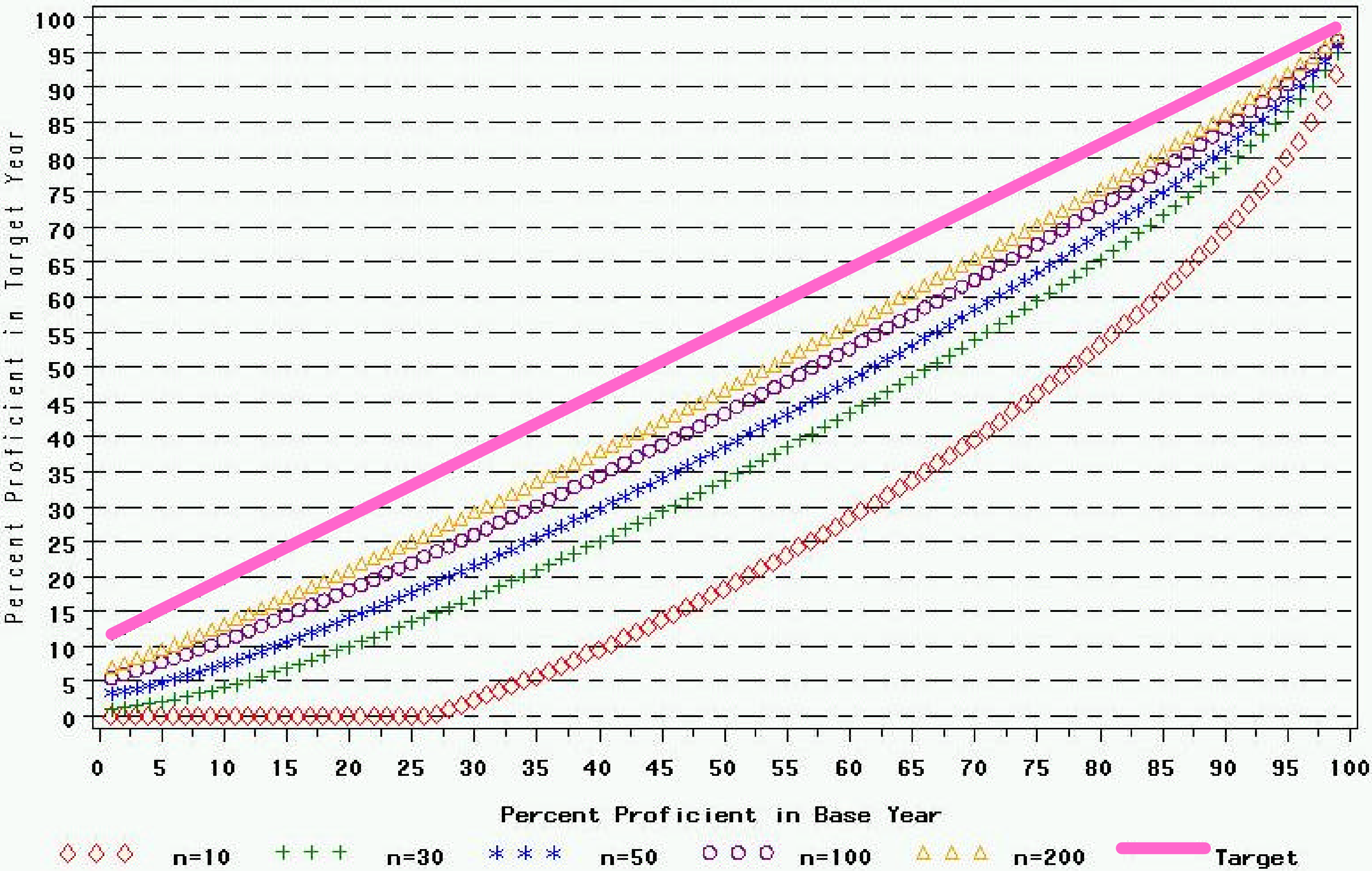
NCLB: Determining AYP Through Status
Relationship between Required Status Target and 95% Confidence Interval
Required Percentage of Students
Status Target
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
n=10
n=30
n=50
n=100
n=200
Target

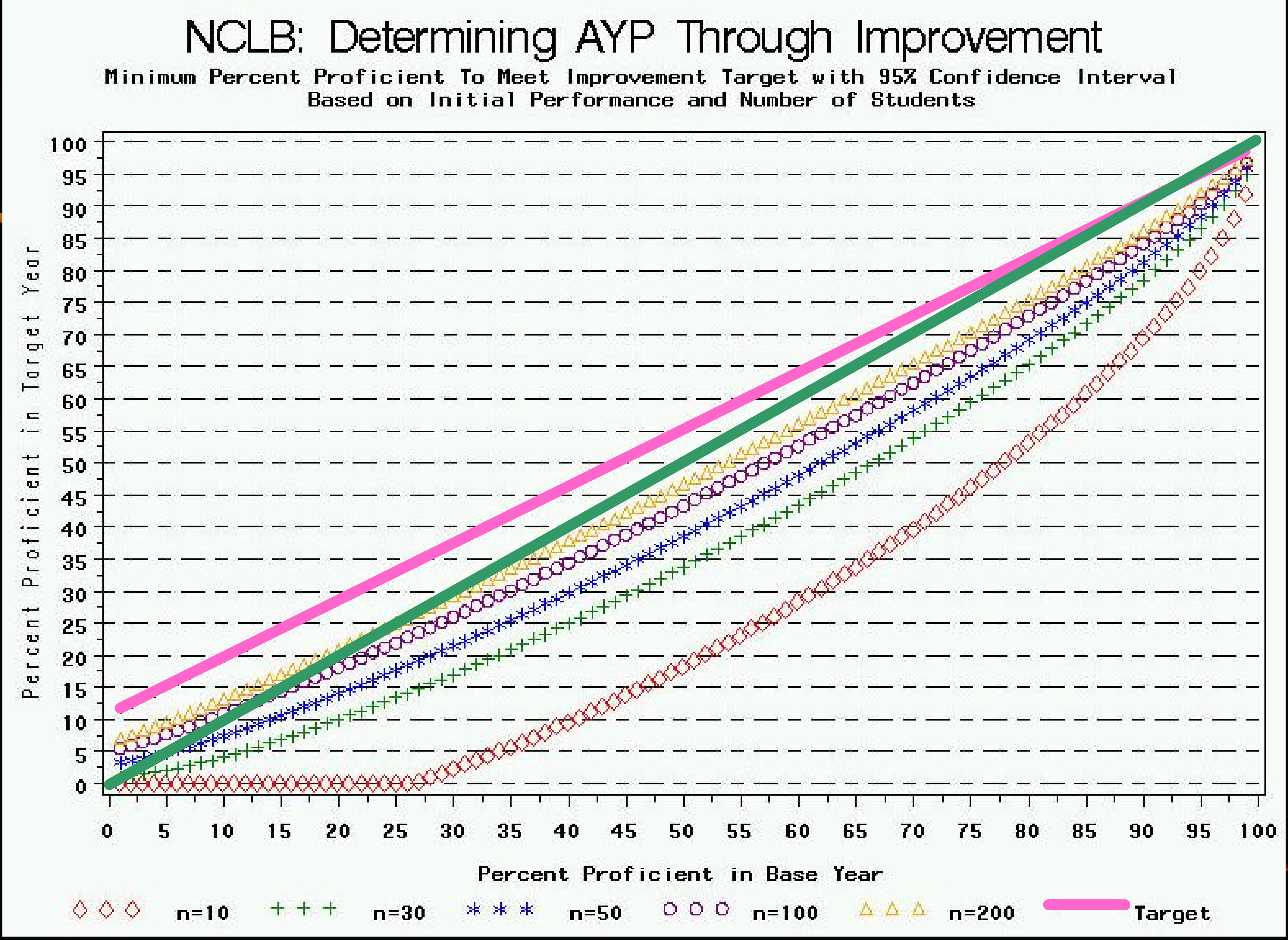


NCLB: Determining AYP Through Improvement
Relationship between Required '10% Improvement' and 95% Confidence Interval
Percentage Points
0
5
10
15
20
25
30
35
40
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Percent Proficient in Base Year
n=10
n=30
n=50
n=100
n=200
Required Gain

NCLB: Determining AYP Through Improvement

Minimum Percent Proficient To Meet Improvement Target with 95% Confidence Interval
Based on Initial Performance and Number of Students




NCLB: Determining AYP Through Improvement
Minimum Percent Proficient To Meet Improvement Target with 95% Confidence Interval
Based on Initial Performance and Number of Students
Percent Proficient in Target Year
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Percent Proficient in Base Year
n=10
n=30
n=50
n=100
n=200
Target

Confidence Intervals vs. Minimum N

- Acceptable practice is to set a minimum number (typically 30-50) of students in group
- That practice is both unreliable *and* invalid
  - Unreliable because 30-50 students is an insufficient number to detect improvement
  - Invalid because schools are not held accountable for subgroups with, say, 29 students

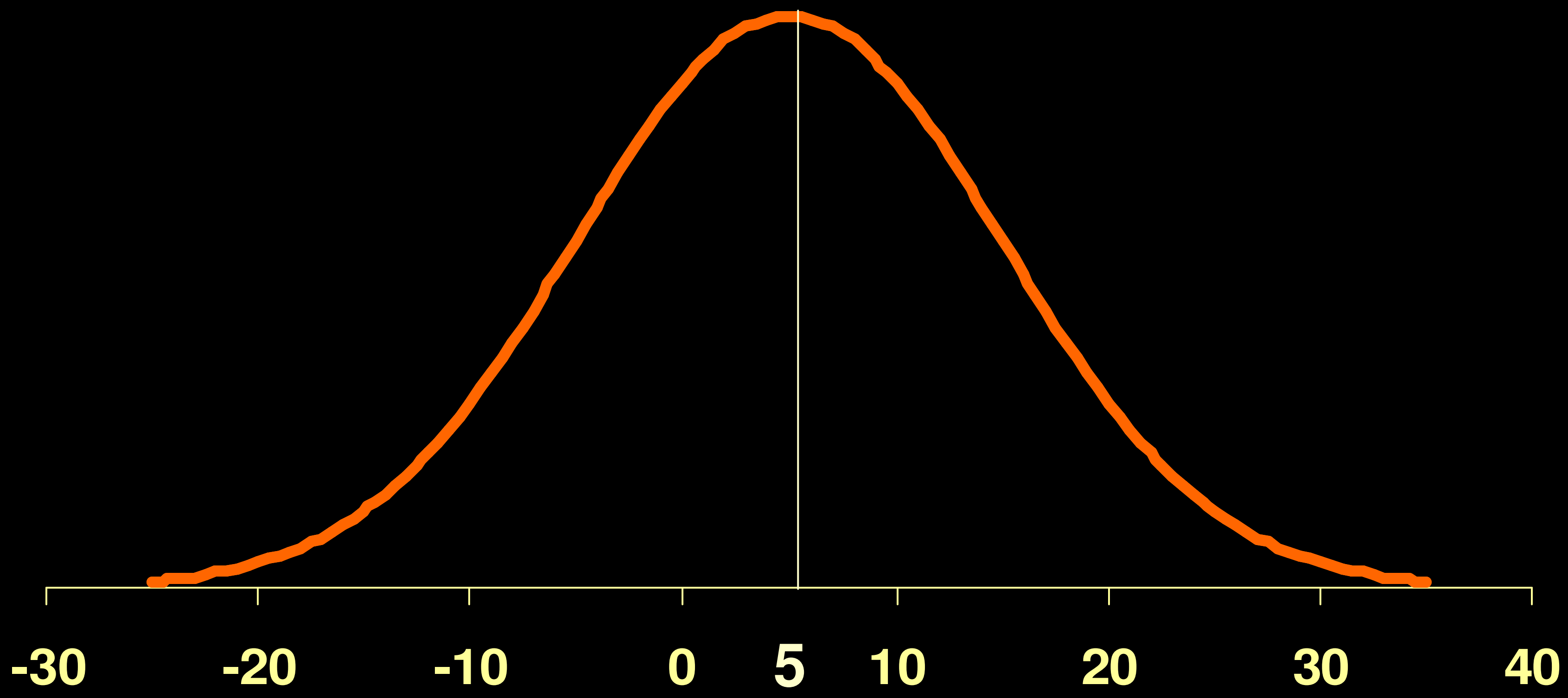
Confidence Intervals vs. Minimum N

- Using confidence intervals for improvement means few schools are identified,but those identifications are reliable
- Using minimum N identifies more schools, but just because you’ve identified *more* doesn’t mean you’ve identified the *right* ones

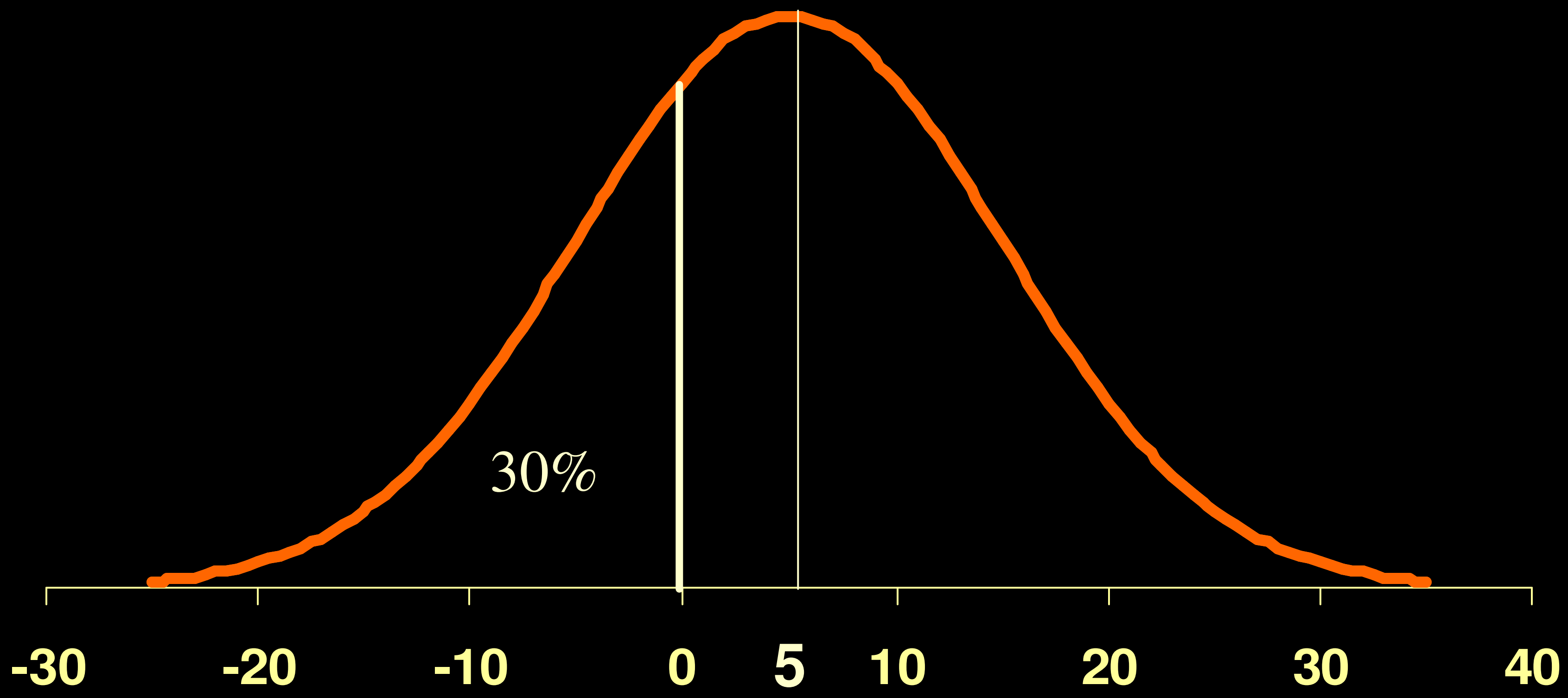
Distribution of Improvement Scores

- If p = .50, groups are required to improve by .05
- If population of school really improves from .50 to .55, what percentage of schools will have observed changes that are 5 percent or more? A *decrease* from previous year?
- What is the bottom 5 percent of that distribution?

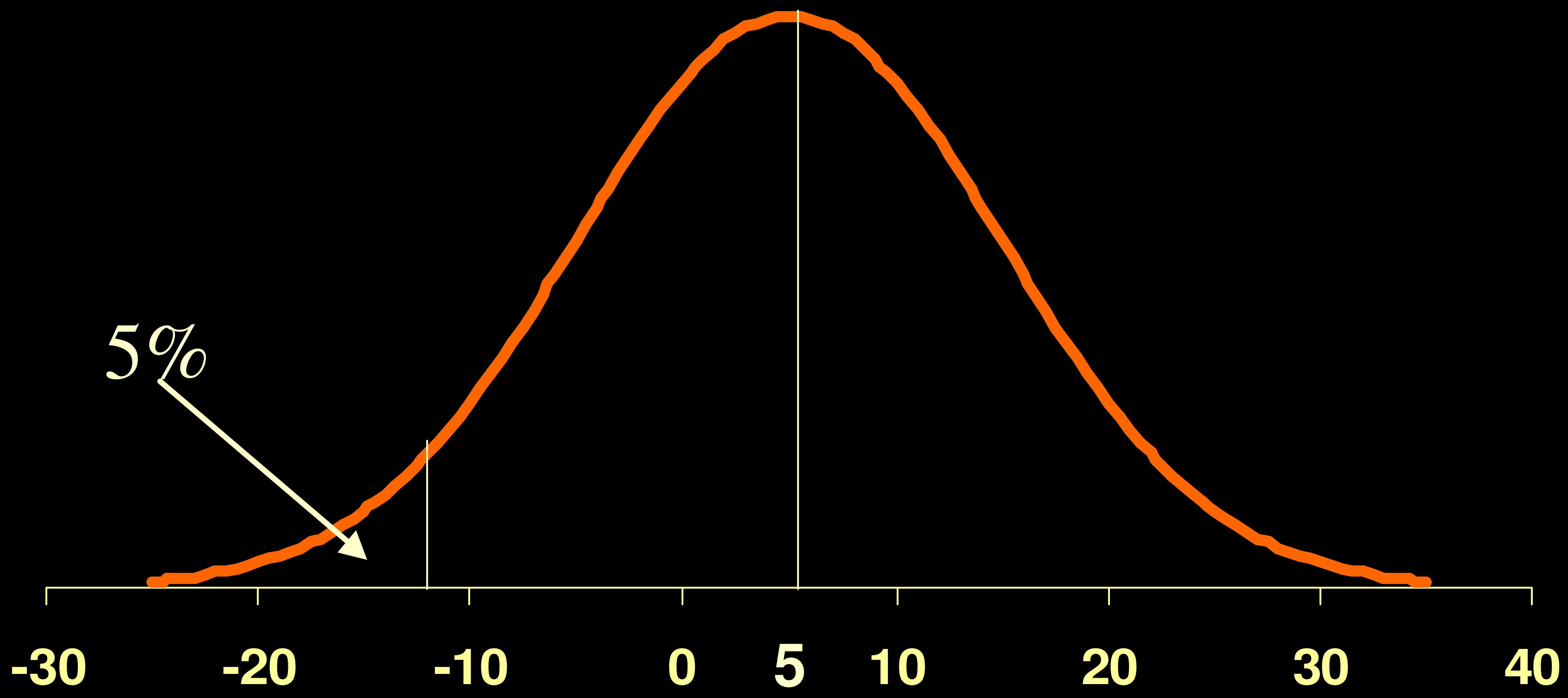
Distribution of Improvement Scores
N = 50, p = .50



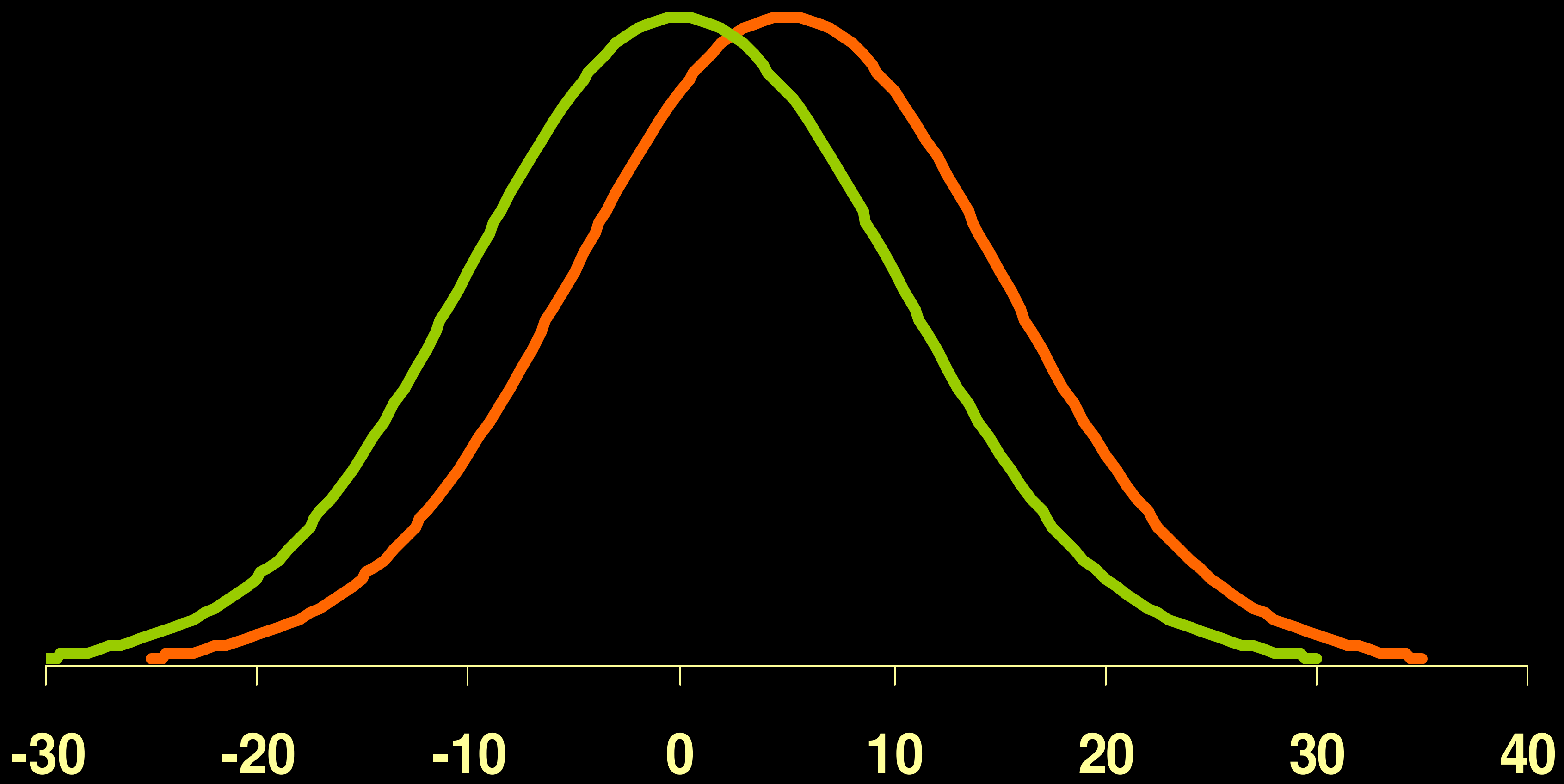
Distribution of Improvement Scores
N = 50, p = .50



Distribution of Improvement Scores
N = 50, p = .50



Distribution of Improvement Scores
N = 50, p = .50



Choosing an *Alpha* Level

- USED guideline is an *alpha* level of .25
- What *alpha* level should be chosen for each subgroup if the desired *alpha* level for the school is .25 (a *school-wise alpha* level of .25)?
- If 18 tests are run, and all are independent, each test needs to be at the .015 level

The Study

- Drew a random sample of 300 students from a state
  - Six subgroups
    - Three ethnic groups
    - Economically disadvantaged
    - Special education
    - Limited English proficient
- Assigned standard scores at random from normal distribution, with mean = 0, sd = 1, to get “Year 1” data

The Study (cont'd)

- Computed percentage Proficient (Proficient was a z-score > 0)
- Computed number of additional students that would be needed for 10 percent reduction in non-proficient for every subgroup
- Changed Not Proficient to Proficient for that number to get "Year 2" data after "improvement"

Summary of Study Data

Group	Number of Students	N and % Proficient	
		"Year 1"	"Year 2"
Whole School	**300**	**150 (50)**	**165 (55)**
Subgroup 1	238	121 (51)	133 (56)
Subgroup 2	105	59 (56)	64 (61)
Subgroup 3	44	26 (59)	28 (64)
Subgroup 4	28	10 (36)	12 (43)
Subgroup 5	29	13 (45)	15 (52)
Subgroup 6	22	12 (55)	13 (59)

Summary of Study Data

Number of Subgroups	Number of Students
1	173
2	90
3	35
4	2

Next Step in Study

- Drew 3500 schools of 300 students each, drawing with replacement from the “populations” created
- Computed whether each subgroup and the school as a whole made AYP under different rules
- Keep in mind that every draw was supposed to make AYP—all had reduced non-proficient by 10 percent

Results of Study

Alpha	With Improvement	
	% Subgroups Making AYP	% Schools Making AYP
.50	50	5
.05	95	75
.01	98-99	93

Results of Study

Alpha	With Improvement		No Improvement	
	% Subgroups Making AYP	% Schools Making AYP	% Subgroups Making AYP	% Schools Making AYP
.50	50	5	11-38	1
.05	95	75	66-89	40
.01	98-99	93	86-97	71

Cautions

- This study is *conservative*
  - 7 groups and 1 test vs. 9 groups and 2 tests
- States should run a similar test on their own data to determine what group-level alpha needs to be to have a school-wise alpha rate of .25

Conclusions

- To have a school-wise alpha rate of .25, you need to use an alpha rate of .05 for subgroups
- Given the requirements of NCLB, improvement cannot be measured reliably for most schools
- But NCLB requires that AYP be defined "…in a manner that is statistically valid and reliable."
- So, come to tomorrow's session on longitudinal designs and see at least one way of doing that