The A*Star® Audit

A Review of Standardized Test Administration

A window onto classroom test administration practices

Eliot R. Long
366 Clinton Street, Brooklyn, NY 11231
Tel: 1-718-624-4216

Reliability & Test Item Response Patterns

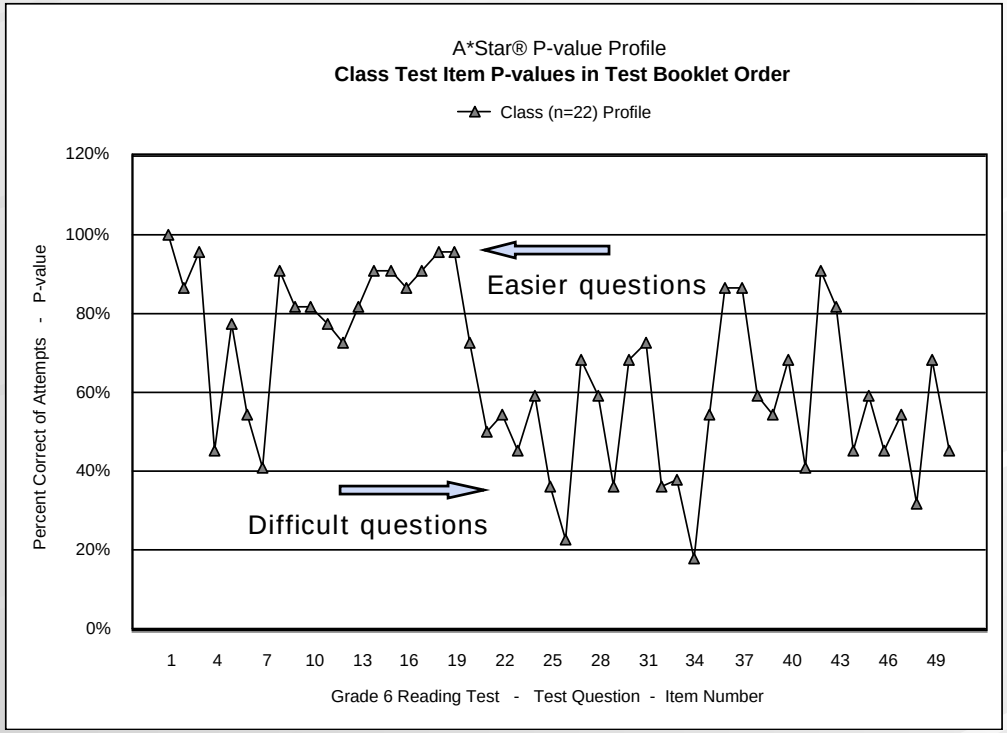
- Test score reliability depends on the consistency of student patterns of success with the patterns of variation in test item difficulty.
- Professionally developed tests provide an environment where students may demonstrate the level of their achievement through a well defined pattern of success.
- When this pattern is broken or distorted, either the test development or the test administration has failed to provide or maintain the necessary environment.

Class Profile of Test Item Success

Individually, students may make a few lucky guesses, careless mistakes, or otherwise unexpected responses. As a class, these variations balance out and a stable pattern of achievement emerges.

The pattern of success for students as a group may be illustrated by a simple plot of the percent correct at each test question.

For this “Class Profile” the p-values are plotted in the same order that the questions appear in the test booklet.



Class Comparison to its Skill Level Norm

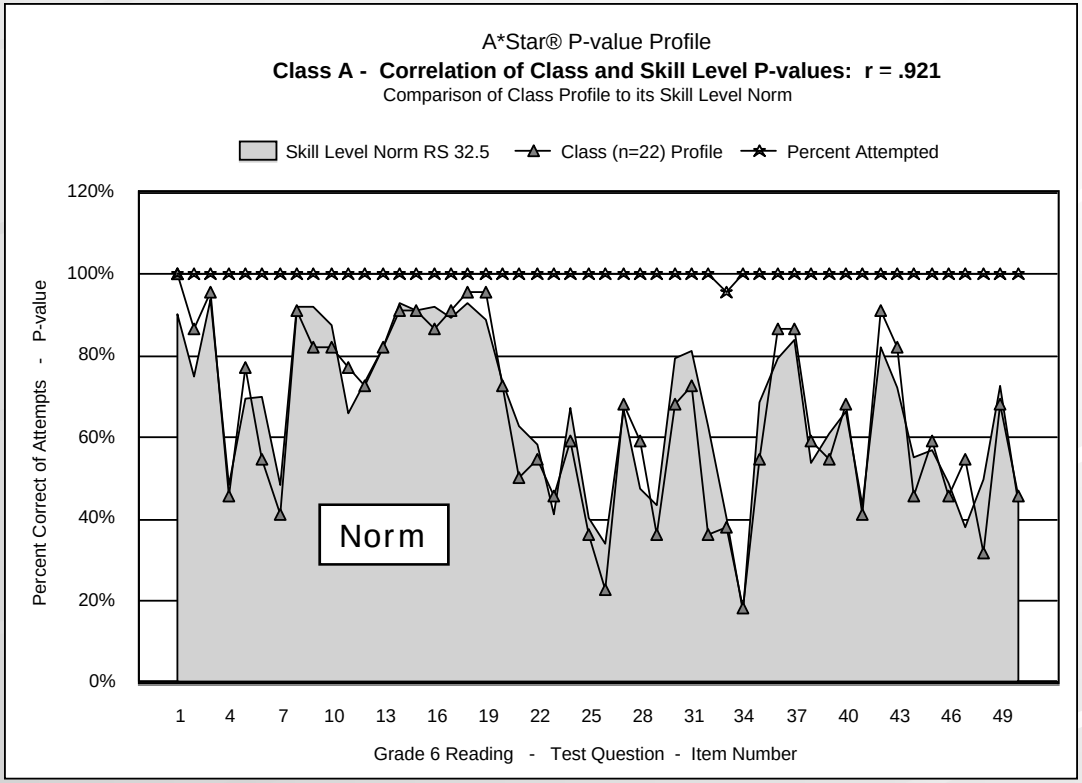
Interpretation of the class profile is supported by
a comparison to an appropriate norm profile

The Norm - The norm is created by grouping all classes at the same class average score. The norm p-values are represented by the upper margin of the shaded area.

P-value Correlation

We may evaluate the comparison of the class to its norm in many ways. One example is by a correlation of the class and norm p-values. The correlation in this example is:

n = 50; r = .921



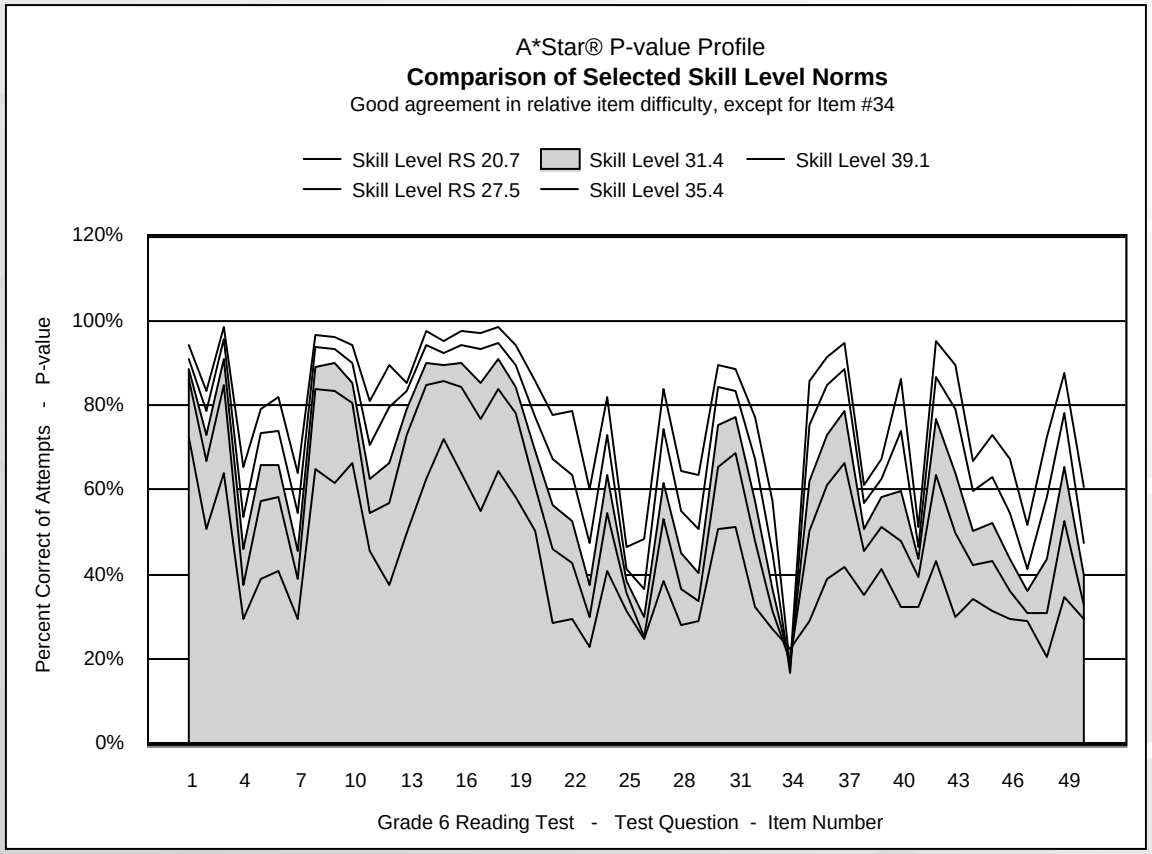
Norms Set by Peer Groups

A range of norm profiles represent a range of skill levels

Classes at different class average score levels are grouped to provide a range of skill level norms. Each individual class is then evaluated against a norm representative of its peers.

Note that the same pattern of rising and falling p-values is repeated at each skill level, only at generally higher or lower p-values.

Note also the exception for question #34. This item has confusing elements in the item stem and does not contribute to measurement.

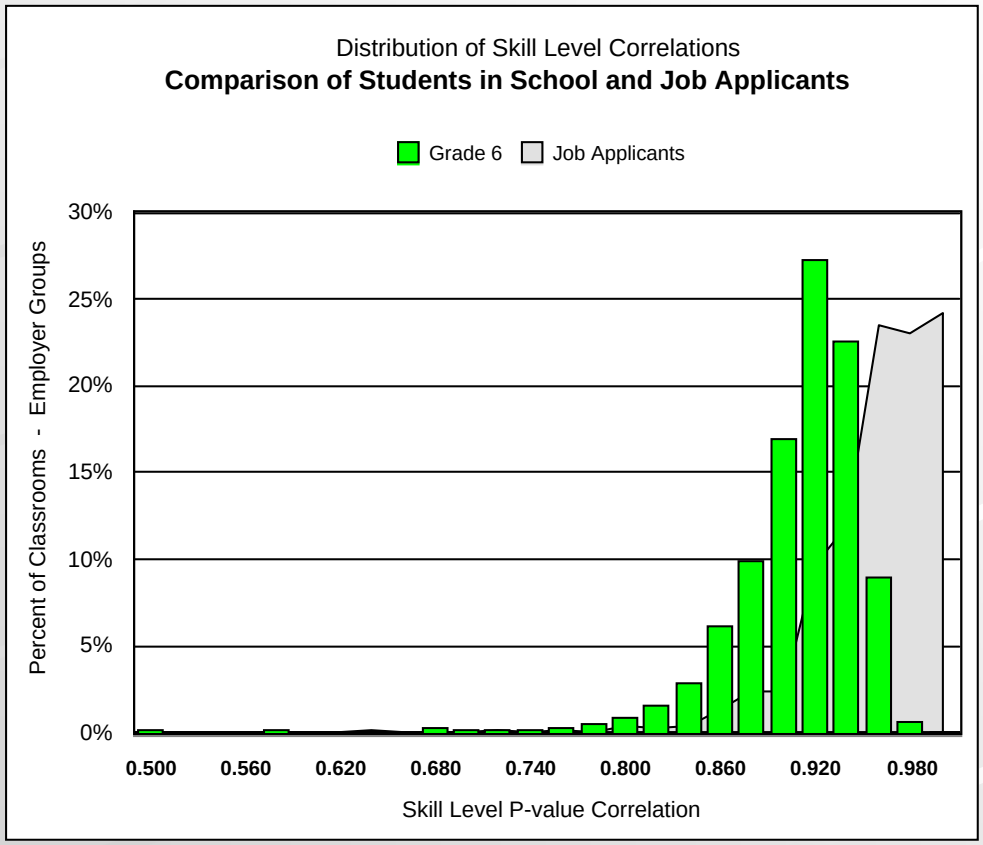


Consistency of Test Administrations

When all classroom profiles are correlated with their appropriate skill level norms, the distribution of correlation coefficients indicates the consistency of the school district test administrations.

A comparison of classroom groups with job applicant groups (tested by employers) indicates lower consistency in classroom test administrations.

Schools median r = .907
Employers median r = .958



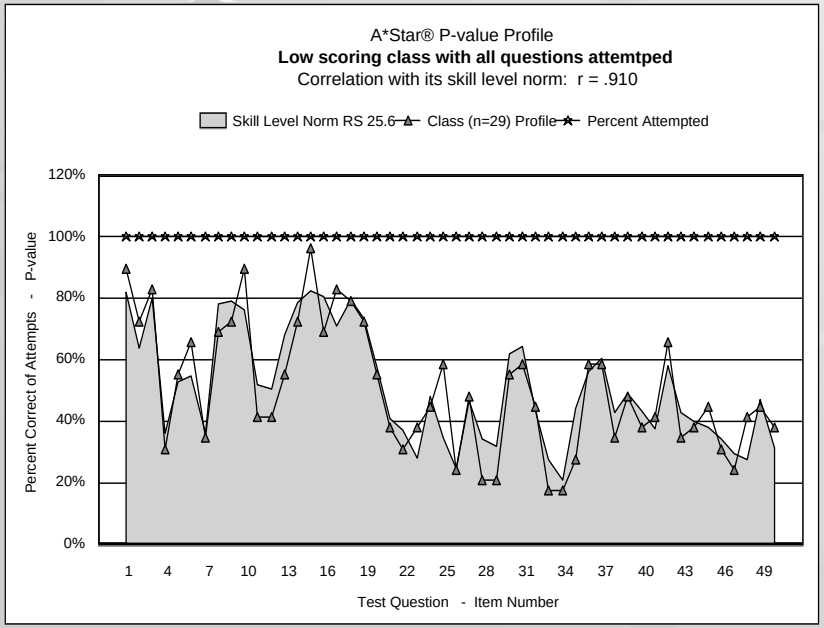
Proctor Effects in a Northeastern School District

Unexpected High Volume of Test Answers

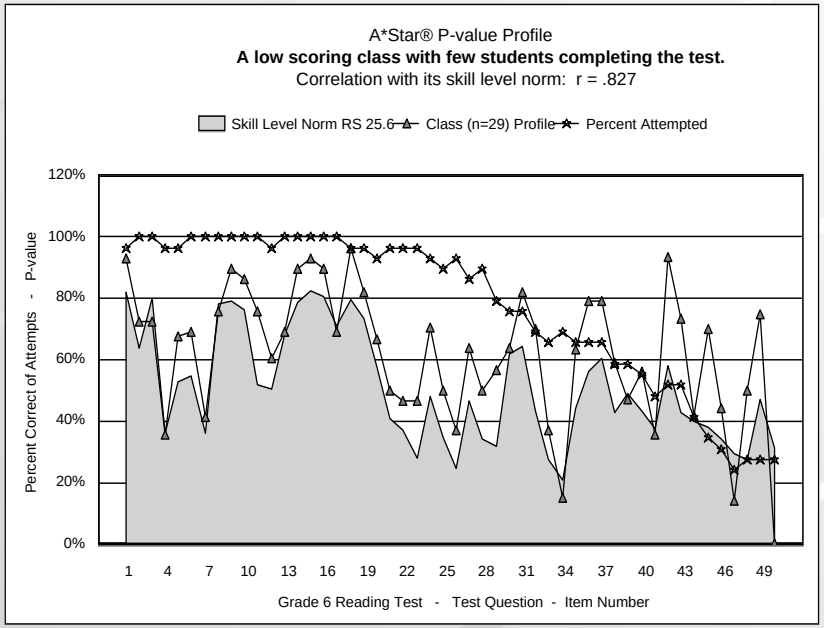
Variations in teacher encouraged guessing to complete all test answers creates non-skill related variation in response patterns.

Two classrooms with the same class average score and yet substantially different student test work behavior. The percent of students who answer each test question is represented by the line with small stars.

Correlation with the norm: r = .910



Correlation with the norm: r = .827

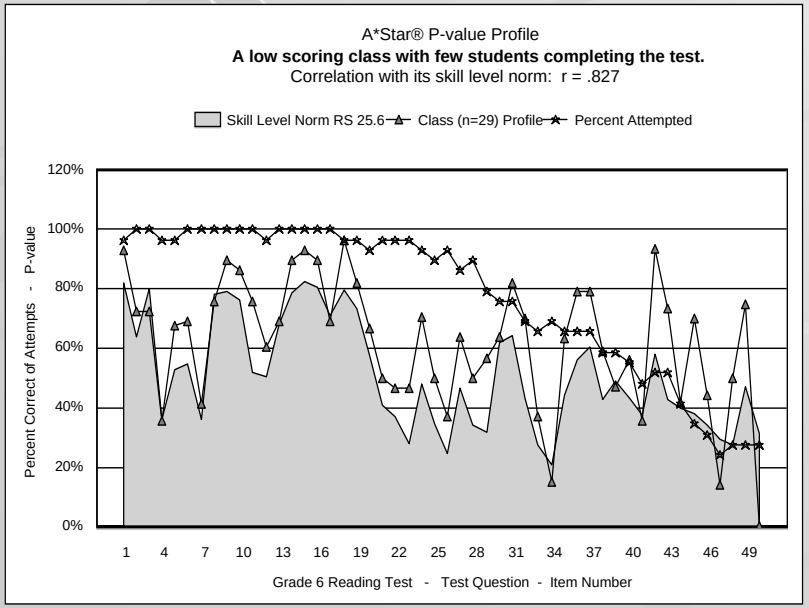


Added Random Guessing Improves Agreement with the Norm

When random guessing is added to replace answers left blank, the class average score is raised by 11% and the correlation with the skill level norm is raised from .827 to .906.

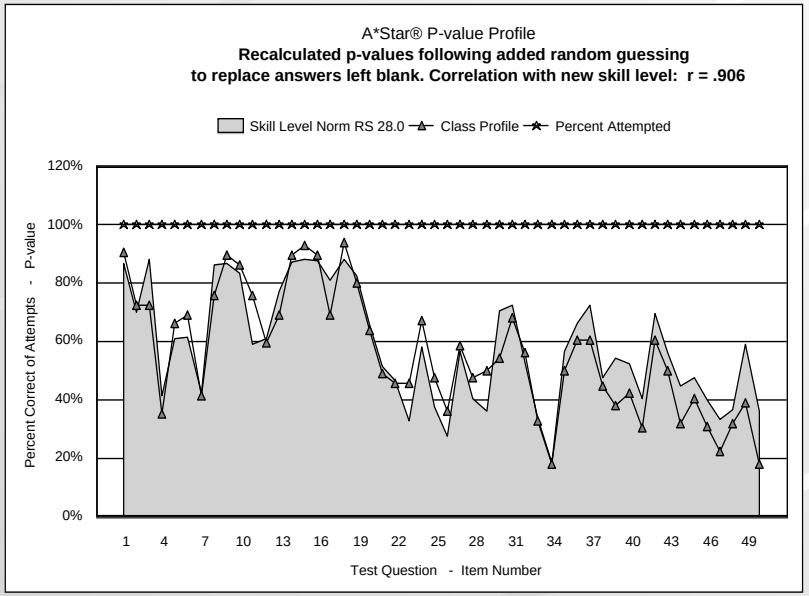
Test results as the test was administered.

Correlation with the norm: r = .827



Results after adding 1/4 correct for each answer left blank.

Correlation with the norm: r = .906



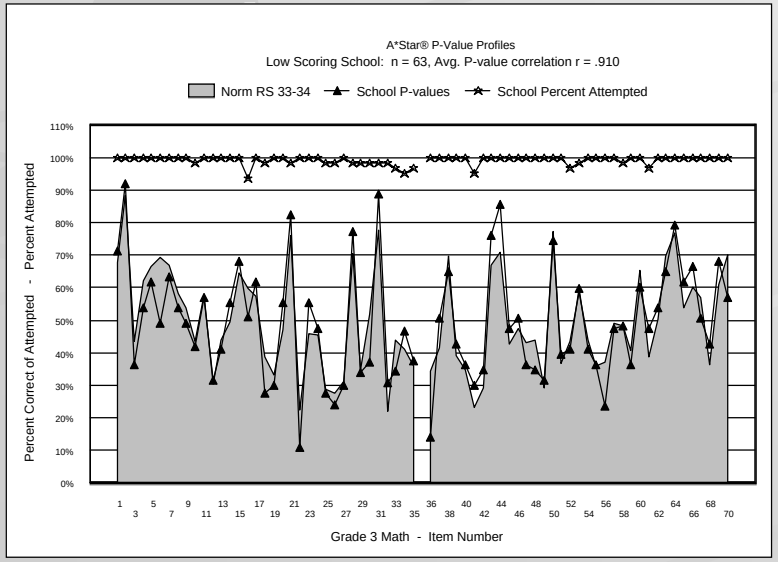
Proctor Effects in the Midwest

Unexpected, High Volume of Answers

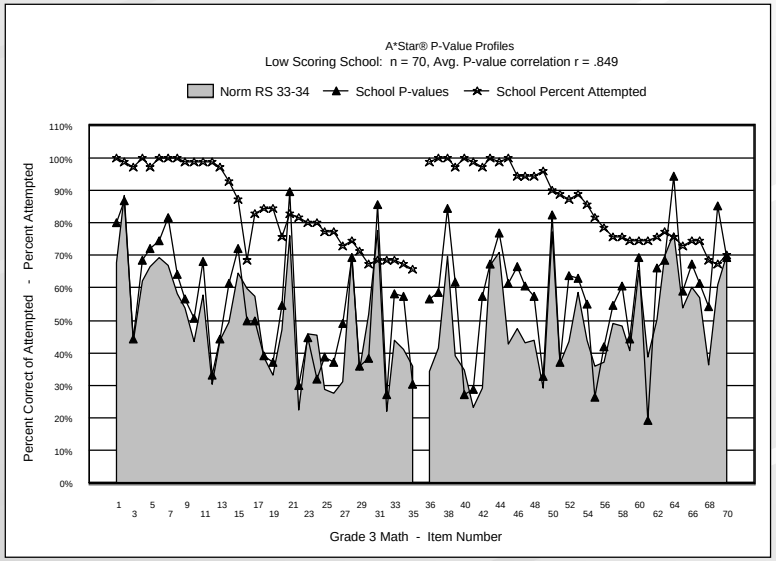
From elementary schools in the Midwest – a grade 3 math test. Two schools with the same school average test score, one school with high test completion, one with substantially lower completion.

The test was comprised of two 35 item test booklets, administered in two sessions, one in the morning and one in the afternoon.

Correlation with the norm: r = .910



Correlation with the norm: r = .849



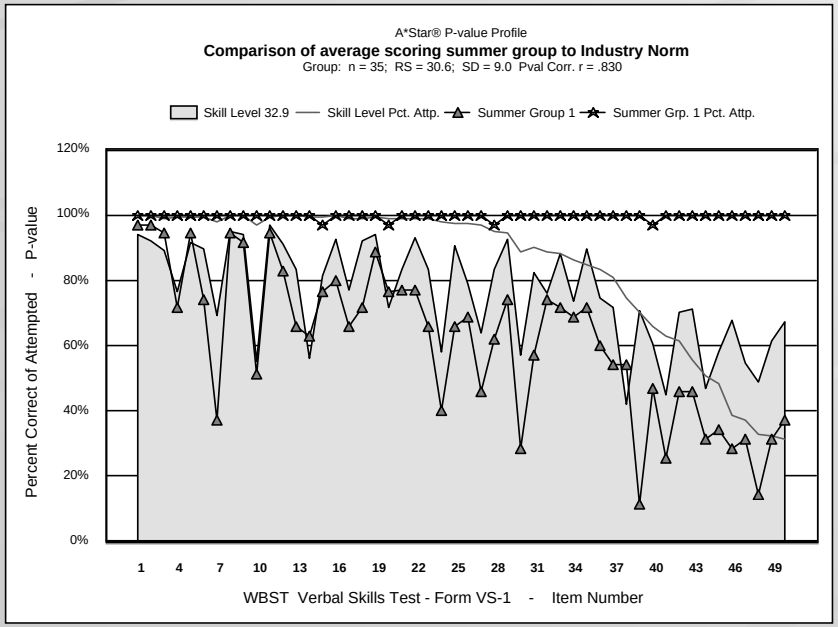
Proctor Effects in the Mid-Atlantic

Unexpected, High Volume of Answers

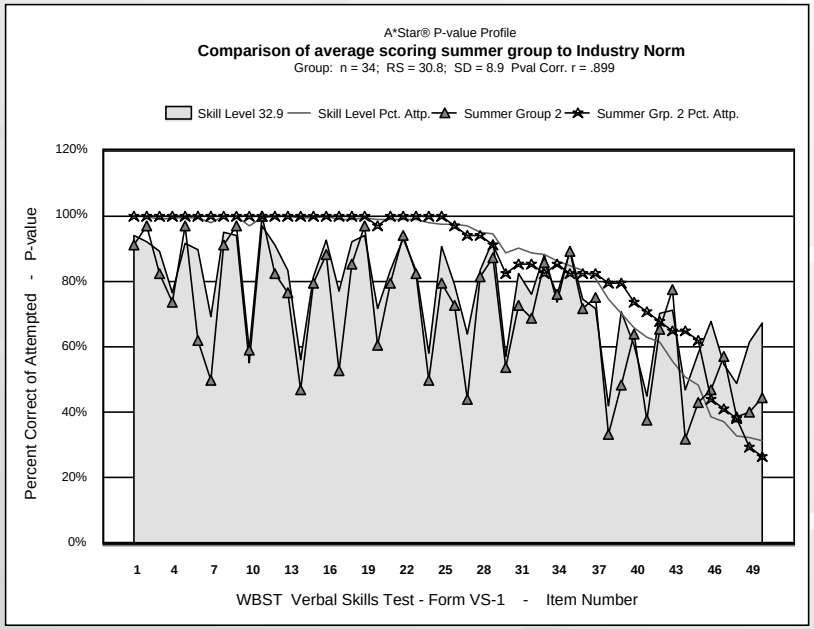
Testing for a special, federally funded, summer program in a Mid-Atlantic State
Grades 6 through 10

Test: A basic verbal skills, pre-employment test
Skill Level Norms: Based on job applicants
Norms do not include encouraged guessing.

Correlation with the norm: r = .830



Correlation with the norm: r = .899



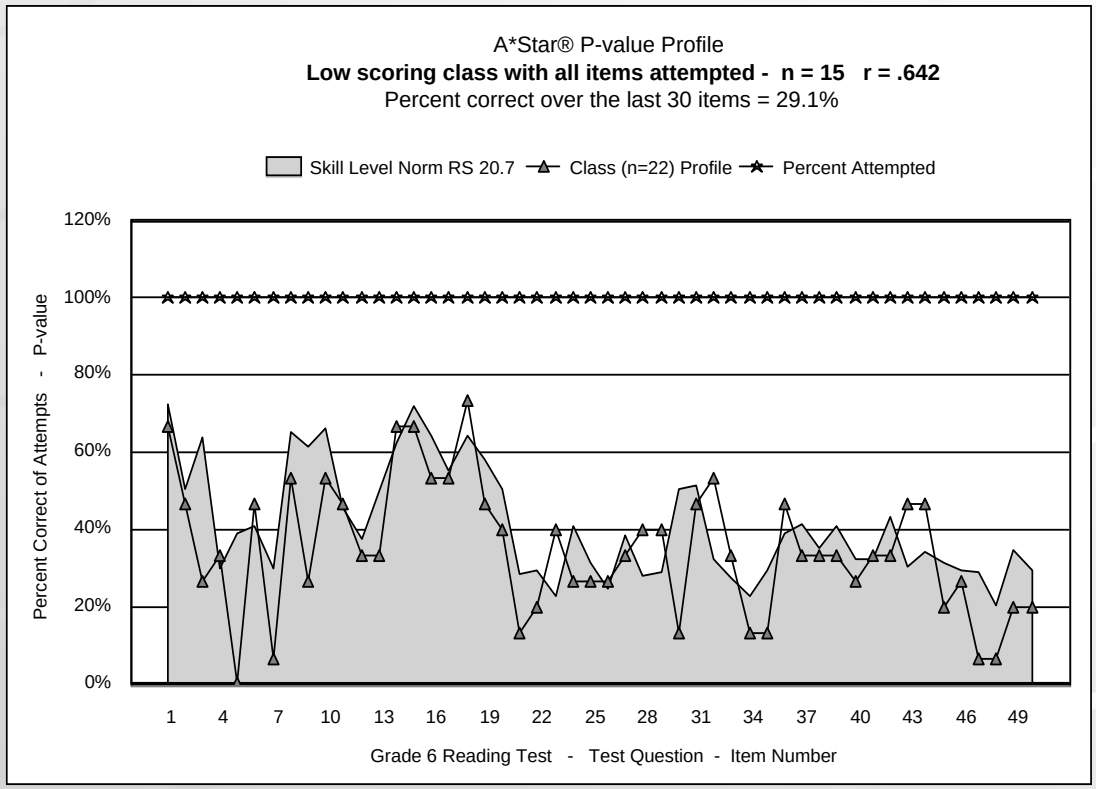
High Volume of Guessing
In Low Performing Classrooms

In low performing classrooms, random guessing may make up more than 50% of test answers.

In this classroom, the average p-value for questions 21 to 50 is 29.1%

The teacher influence in this classroom is likely to have rushed students to guess where taking more time would have been more successful. The low correlation with the norm suggests confusion.

Correlation with the norm: r = .642

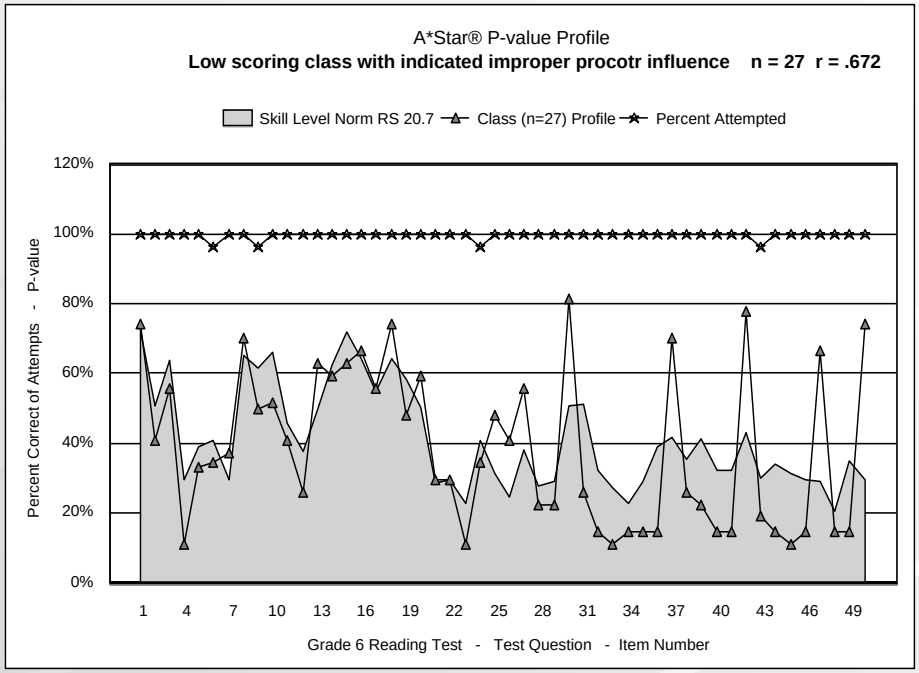


Indications of Improper Influence

Teacher encouragement to guess may include advice on guessing strategy and may lead to inadvertent or overt assistance with correct answers. This assistance may take many forms, for example:

- **Guessing strategy**; i.e. pick a middle size answer on math tests.
- **Brief instruction**; i.e. remember, to divide fractions, invert and multiply.
- **Hints & direct answers**.

Significant, improper influence is revealed by improbable breaks in the class response pattern.



Educator's rationalization:

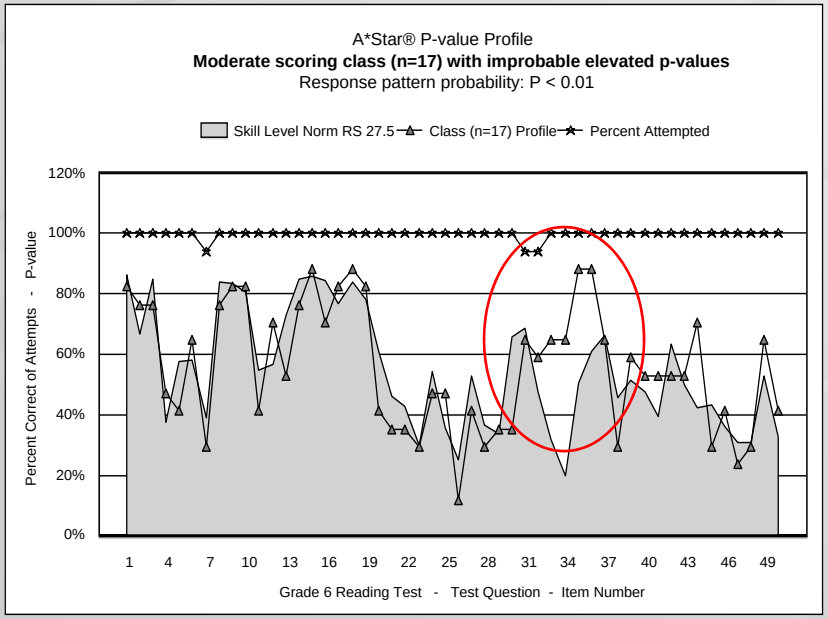
"when a teacher 'helps' a struggling student with one or two challenging test items, we may view it as a small and forgivable infraction when compared to the potential motivational and psychological costs of that student failing yet another test."

Collateral Damage by S. L. Nichols and D. C. Berliner, Harvard Education Press 2007, p34

Improper Proctor Influence

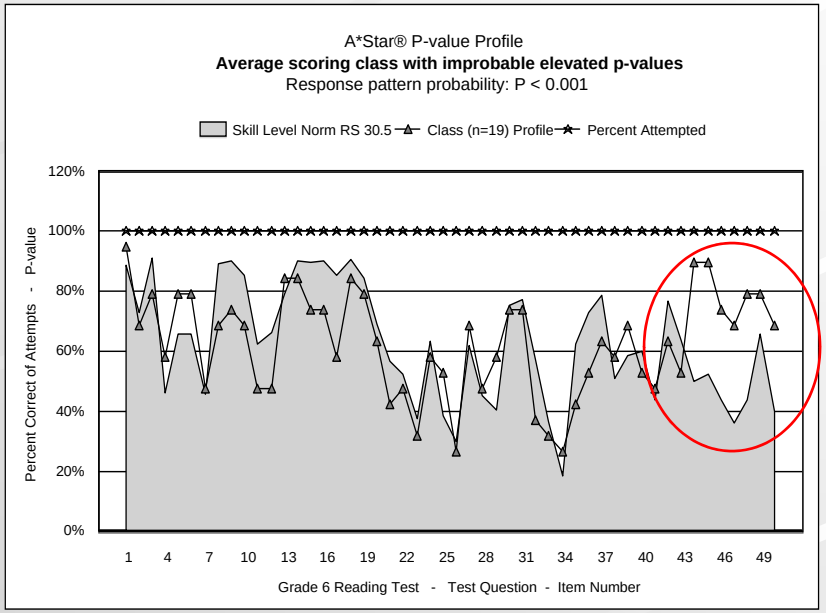
Proctor influence ranges from positive to moderately negative to a serious undermining of the assessment. Significant improper influence leads to measurable deviations in classroom response patterns.

Response pattern probability: P < 0.01
Correlation with the norm: r = .713



Likely assistance with a confusing test section

Response pattern probability: P < 0.001
Correlation with the norm: .579



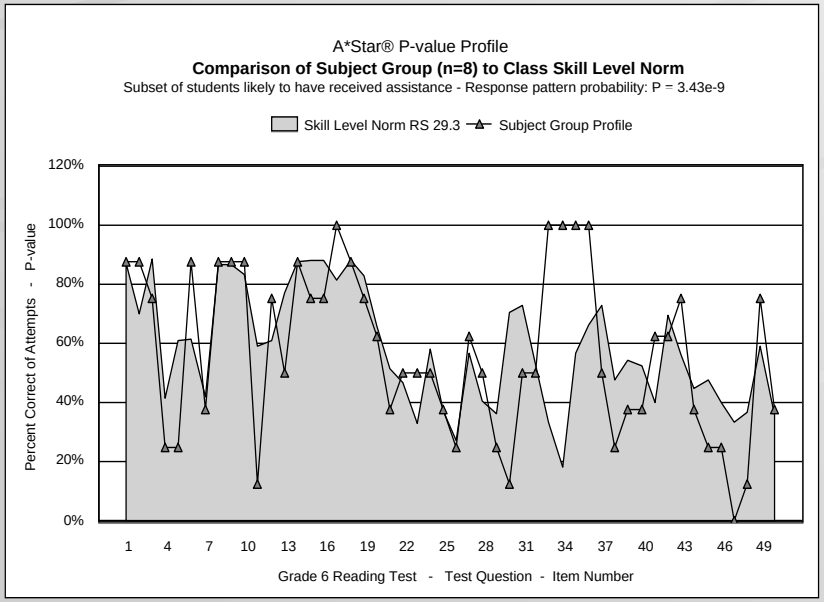
Likely addition of correct answers after testing.

Subject Group Analysis

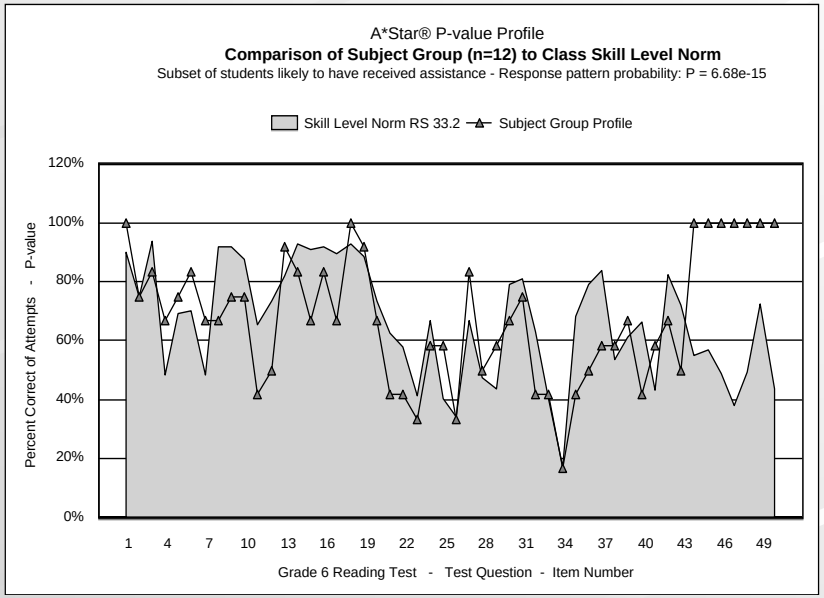
Identify those most likely subject to improper influence

Most often, improper teacher influence is unplanned and disorganized.
Yet, where the influence is persistent, subsets of students will be identified with matching, unlikely response patterns.

Subject Group: n = 8 of 17
Response pattern probability P = 3.43e-9
Correlation with the norm: r = .498



Subject Group: n = 12 of 19
Response pattern probability: P = 6.68e-15
Correlation with the norm: r = .342



Encouraged Guessing
Encourages Improper Influence

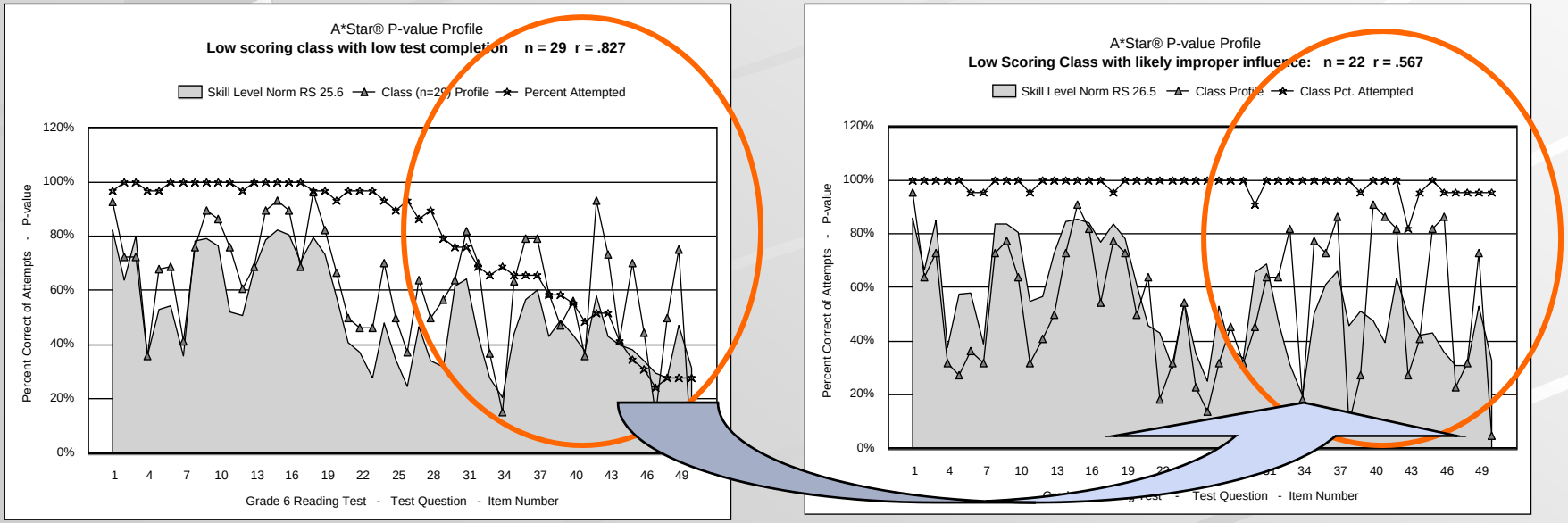
Identified patterns of improper influence most often arise in the area where students would otherwise guess or leave answers blank.

<u>Class without improper influence.</u>

Note higher performance than the norm over early test items. Seemingly erratic p-values at the end are due the fewer, higher achieving students completing the test.

<u>Class with likely improper influence.</u>

Note lower performance than the norm over early test items. Erratic p-values at the end involve most or all students, unlikely without proctor involvement.



Characteristic Patterns
Indicate the Nature of Proctor Influence

Higher than expected, yet variable, performance over a limited test section
- Often indicates proctor explanation or instruction regarding test content.

Higher than expected performance over the later half of test items, yet consistent with the norm pattern of item difficulty
- Often indicates extension of test time limits.

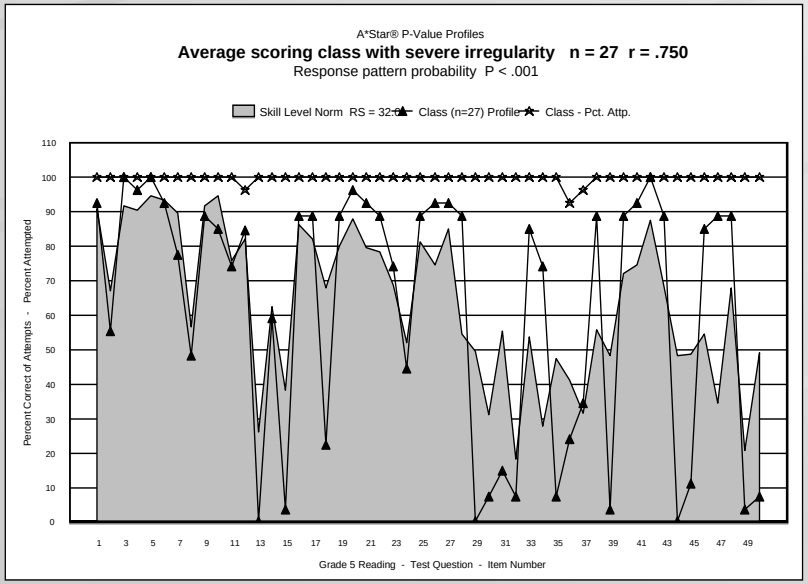
Higher than expected, uniform performance over the later test items
- Often indicates adding correct answers following the test administration. The added correct answers may be limited to answers left blank by lower performing students – resulting in no erasures.

Higher performance over easier questions and lower performance over difficult questions
- Often indicates students copying answers – achieving the same answers right and the same answers wrong. Sometimes the result of seating students close together. Students may be seated in multiple clusters – resulting in greater variation in the copying and in the irregularity of the test response pattern.

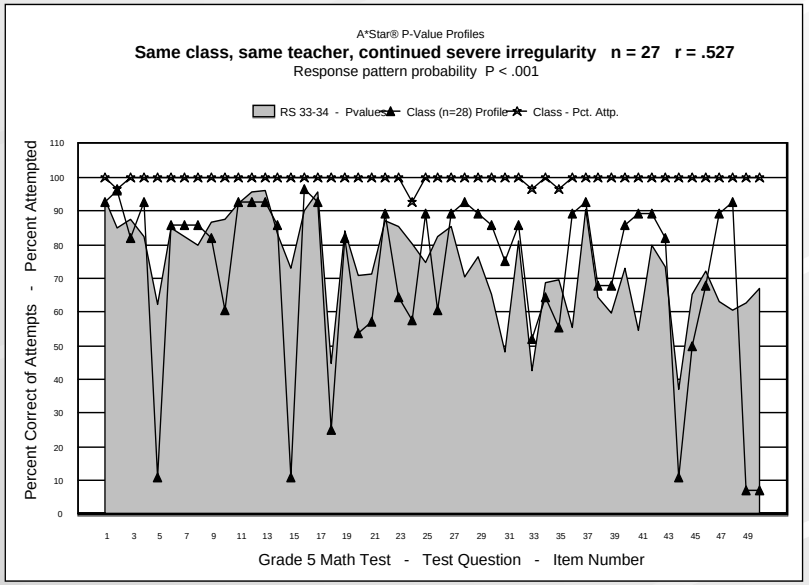
Dramatic deviation from the norm over the length of the test
- Indicates pervasive proctor control over students' responses.

- When No One is Looking -
Same Teacher - Successive Class Response Patterns With Severe Irregularity

April 1st Year
Reading Test
r = .750



May 1st Year
Math Test
r = .527



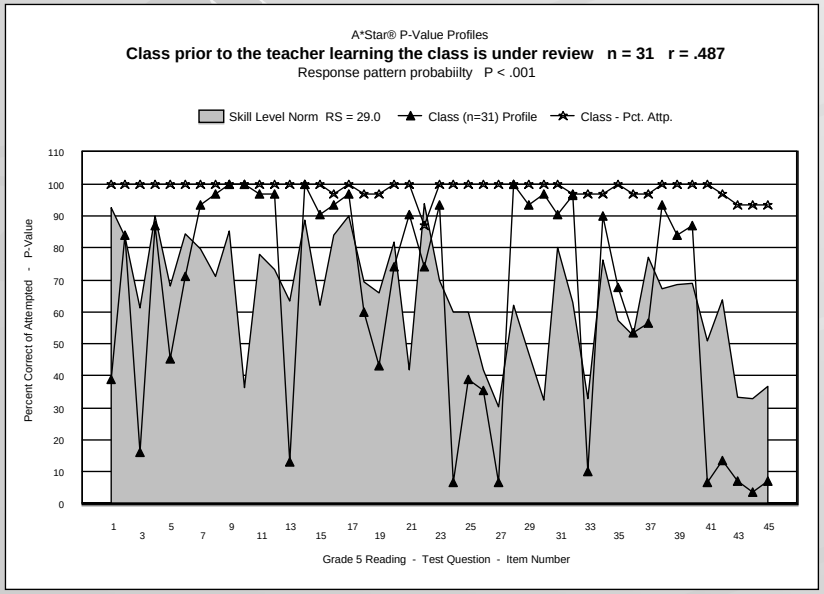
- When Someone is Looking –
Change in Class Response Pattern
Same Teacher After Learning the Class is Under Review

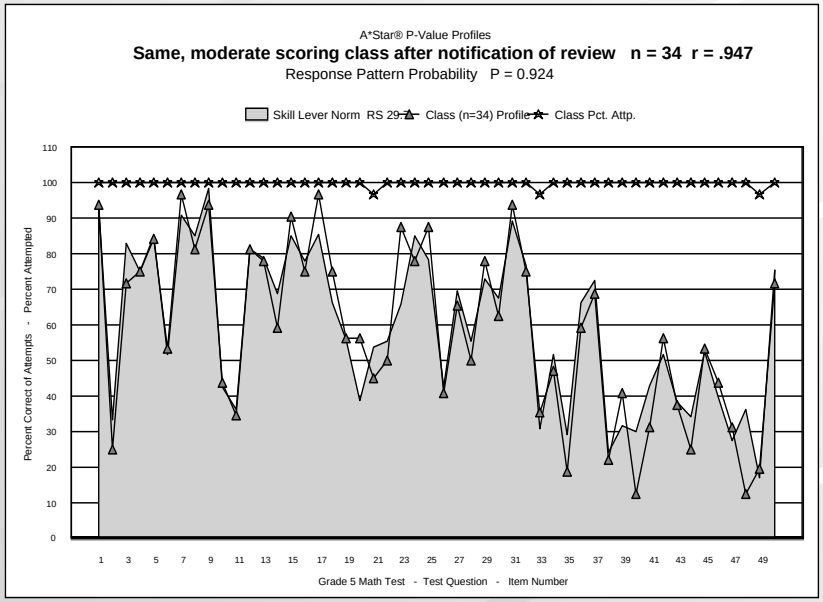
April 2nd Year
Reading Test
r = .487

Prior to notification that class test results are under investigation

May 2nd Year
Math Test
r = .947

Two weeks following notification that class test results are under review





Prevention of Improper Influence in Test Administration

Clear, written test administration policy & procedures

- Specify do's and don'ts with regard to test prep and teacher influence on students' test work behavior.
- Most school district policies only cover non-test related behaviors (i.e. disruptive behavior, bathroom requests, etc.).

Systematic, proactive audits to evaluate test results

- Systematic audits of test results alert teachers/proctors that the results of their test administrations will be evaluated.
- Most school districts have limited, reactive reviews of test results; analyses based on erasures and unusual gains are inconclusive.

Aggressive, fair, follow-up on indications of impropriety

- Investigation of significantly irregular test results to determine the cause and appropriate remedial action is necessary to support the school district policy.

The A*Star® Audit

A comprehensive review of all classroom groups

- Measures each class according to the normal variation of its peers
- Provides a detailed analysis each irregular group

Conducted quickly, independently, without interruption of school activities.

- Audit based on collected test results
- Audit processed by sophisticated, patented, computer software

Audit results evaluate individual groups and overall assessment quality

- Individual classrooms are proactively identified where problems exist
- A searchable electronic file supports school district administration inquiries.