Technology in Testing: Application and Innovation
February 2003

Identifying Rogue Test Administrators

Eliot Long
A*Star®Audits

The Association of Test Publishers

Variations in Test Administration

- We can identify ‘Rogue’ Test Administrators

  Analysis of group test item response patterns identifies a broad range of improper test administrations – including cases of ‘rogue’ test administrators.

- Good Reasons to Make the Effort

  Variations in test administration undermine test reliability and create distrust in the assessment process.

- Management Issues

  Management versus thousands of test prep books and web sites; who wins the proctors’ allegiance? Pressures of accountability without oversight will increase the frequency and severity of improper test administrations.

Evaluating Test Administrations

- Traditional methods

  Erasure analysis, evaluation of unusual test score gains, surveys and interviews of teachers, students, others involved in the assessment process.

- Jacob & Levitt: “Catching Cheating Teachers ..”

  Working Paper (www.nber.org)

  Jacob and Levitt have developed a sophisticated algorithm for identifying unusual item response strings associated with large test score gains in elementary school testing.

- A*Star® Audit of test administrations

  Author’s development in response to the 1996 “Ability to Benefit” testing regulations of the U.S. Department of Education.

A*Star® Audit of Test Administrations

- Vocational School ‘Ability-to-Benefit’ Testing
  9,957 groups evaluated at 661 schools nationwide.

- Industry Job Applicant Testing
  1,323 employer groups evaluated nationwide.

- Public School Testing, Grades 3 – 8
  Over 35,000 classroom groups evaluated from a large urban school system in the Northeast and a statewide assessment in the Midwest.

A*Star® Method of Analysis

1. Establish Test Item Response Pattern Norms
   - by Group Skill Level
2. Measure each Group against its Skill Level Norm
3. Measure Group Variation around each Skill Level Norm
4. Determine 'Irregular' Groups – those that differ significantly from their Skill Level Norm
5. Within each Irregular Group, determine the extent of subset 'Subject Groups'
6. Measure the probability of the Subject Group frequency occurring in groups of the same size and test score distribution.

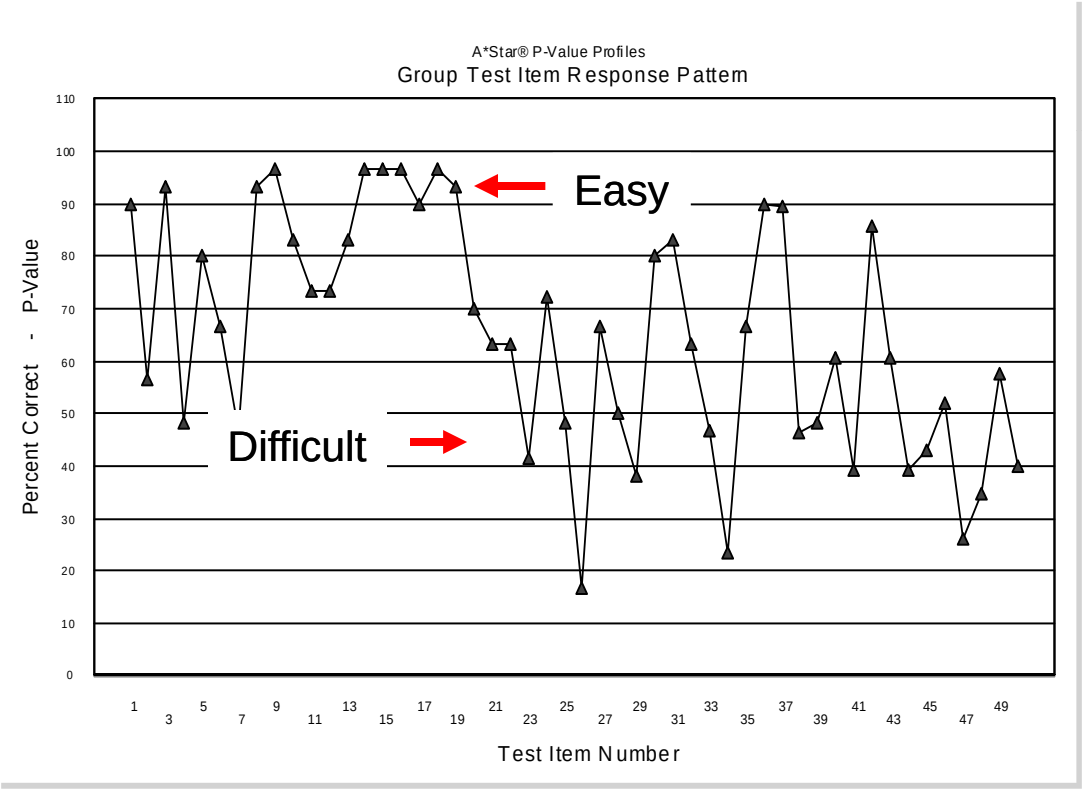
Group / Classroom - Test Item Response Pattern

Percent of Test Takers with Correct Responses

100%

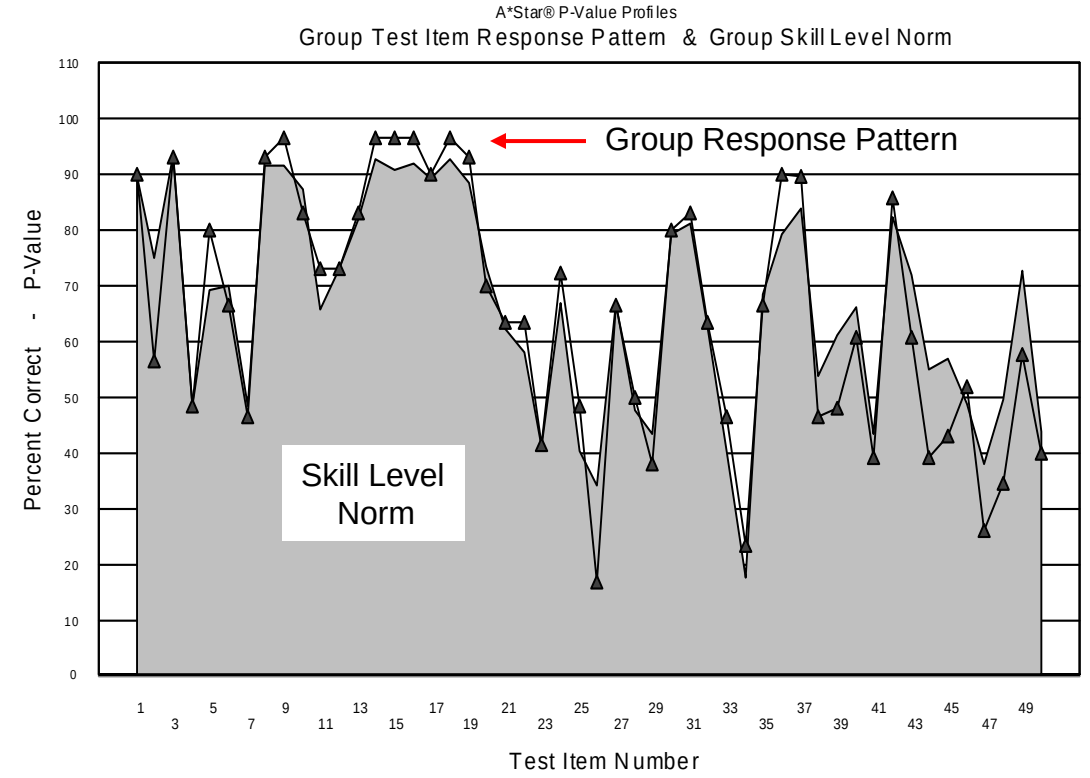
Percent Correct

0%

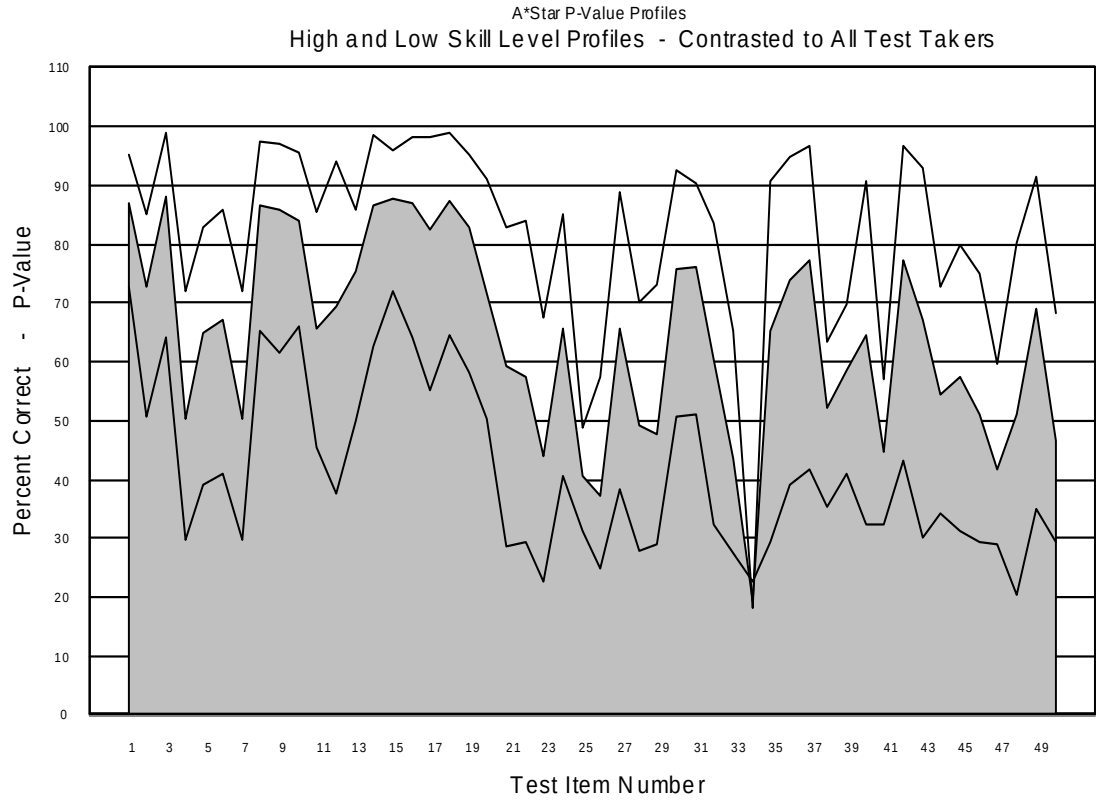


Test Questions in the Order Presented in the Test

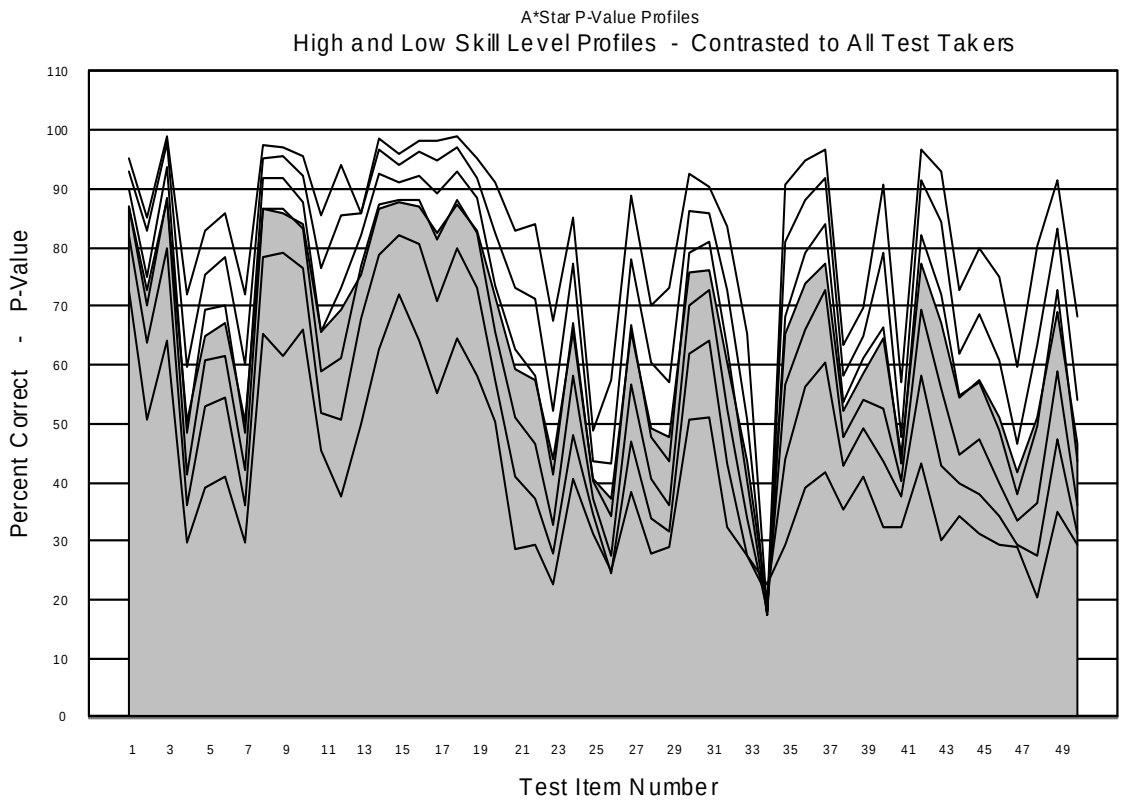
Group Pattern Contrasted To The Group Skill Level Norm



High and Low Skill Levels
Contrasted to the Norm for All Test Takers

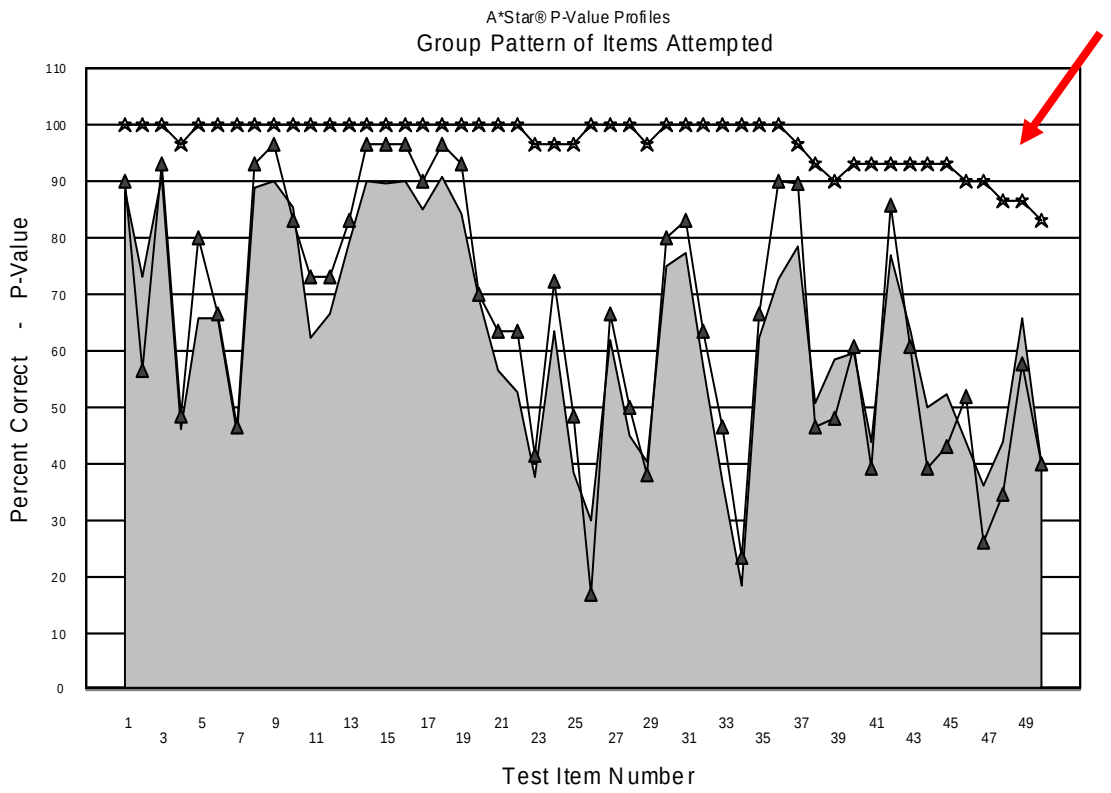


Multiple Skill Levels
Represent the Range of Group Average Scores



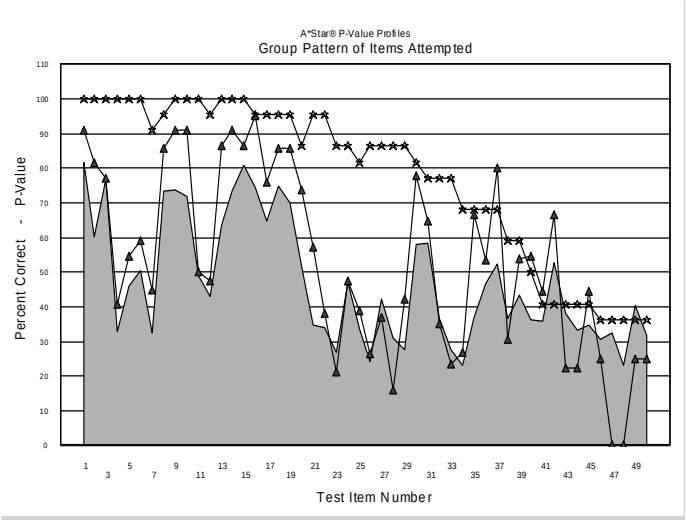
Percent Attempted

Percent of Group Responding to Each Test Item

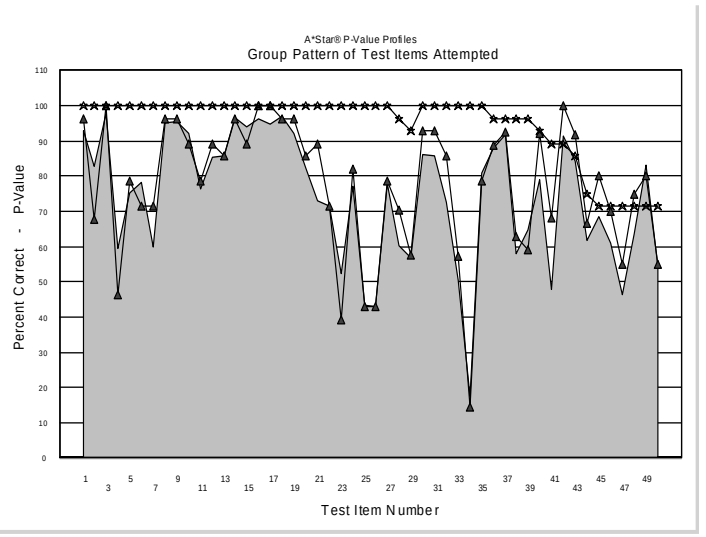


Percent Attempted – Rises with Group Average Score

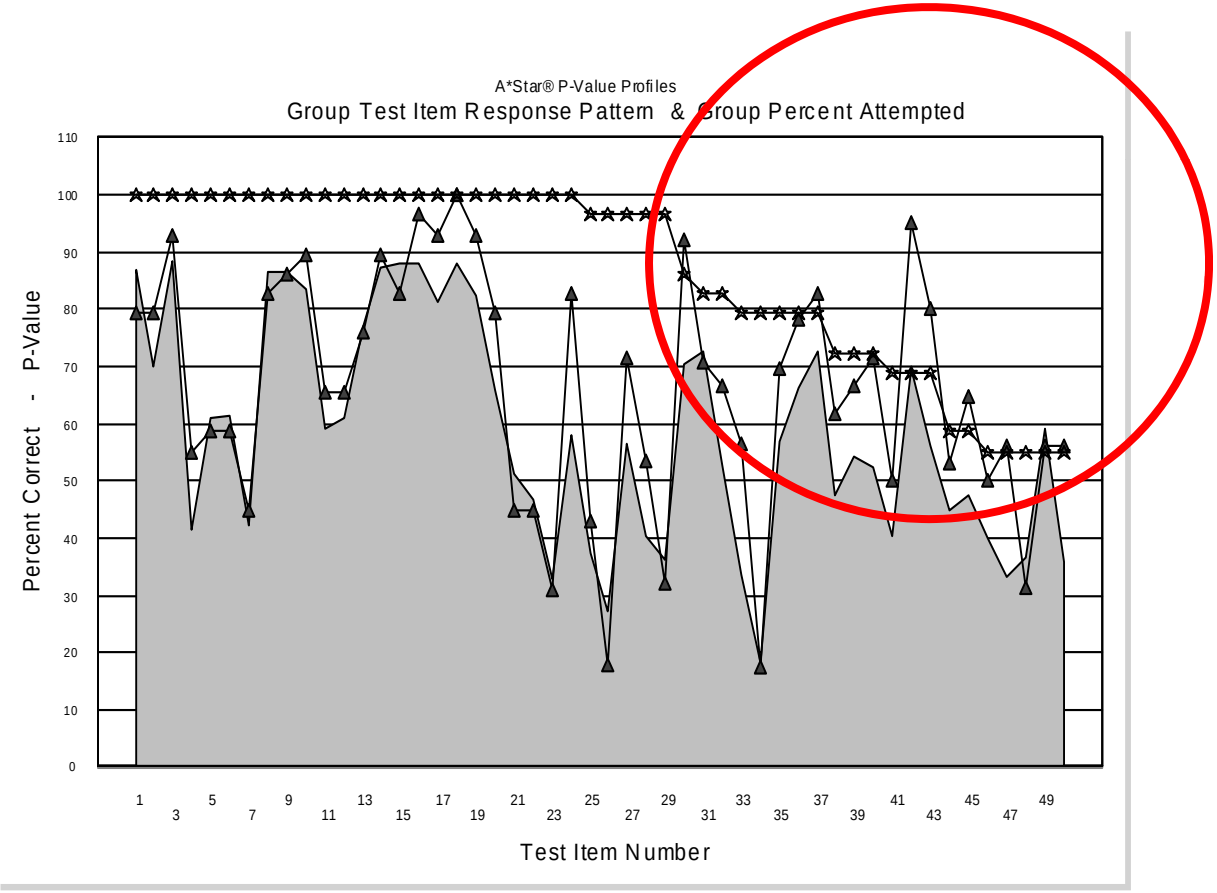
RS = - 1.5 s.d.



RS = +0.75 s.d



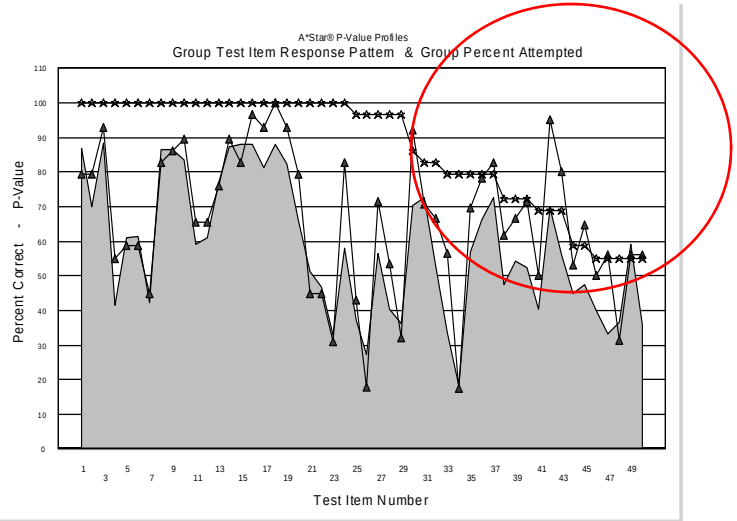
The Proctor's Temptation: How to Fill The Empty Answer Blanks



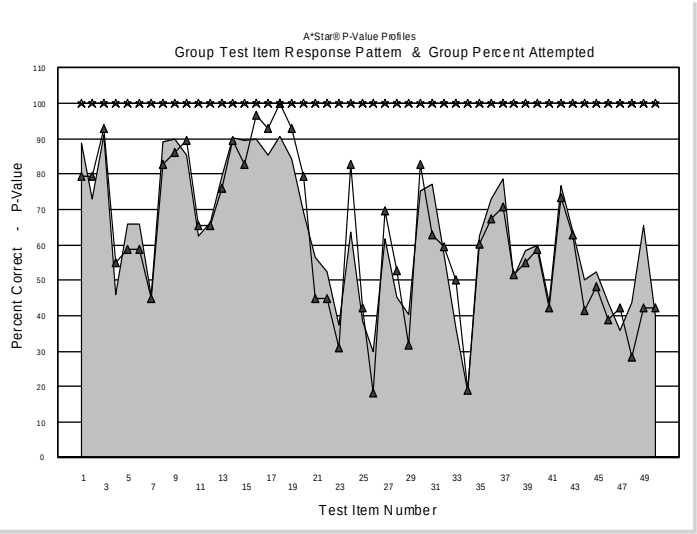
Added Random Guessing
Brings Pattern Closer to the Skill Level Norm

- Added random guessing adds 0.26 s.d. to Group average score
  & 6.2 points in percentile standing among all groups.

Percentile Standing: 31.7



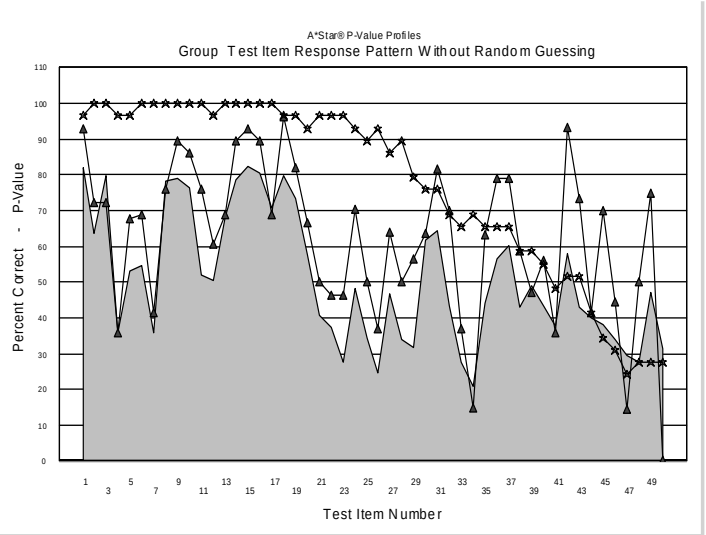
Percentile Standing: 37.9



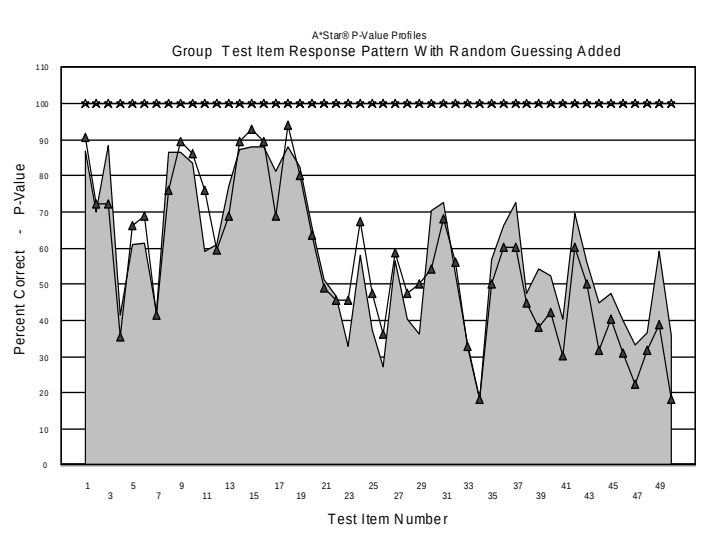
Added Random Guessing Conforms Pattern to Norm & Adds Significantly to Group Average Score

Added random guessing adds 0.45 s.d. to group average score
& 13.3 points in percentile standing.

Percentile Standing: 12.0
P-value Correlation = 0.827



Percentile Standing: 25.3
P-value Correlation = 0.912



Directions for 'Random' Guessing ??

1. Include instructions for guessing with other directions at the beginning of the testing – or wait until the time is about to expire?

2. Address encouragement to guess to the entire test group or to individual test takers?

3. Monitor slower test takers and give individualized encouragement to guess?

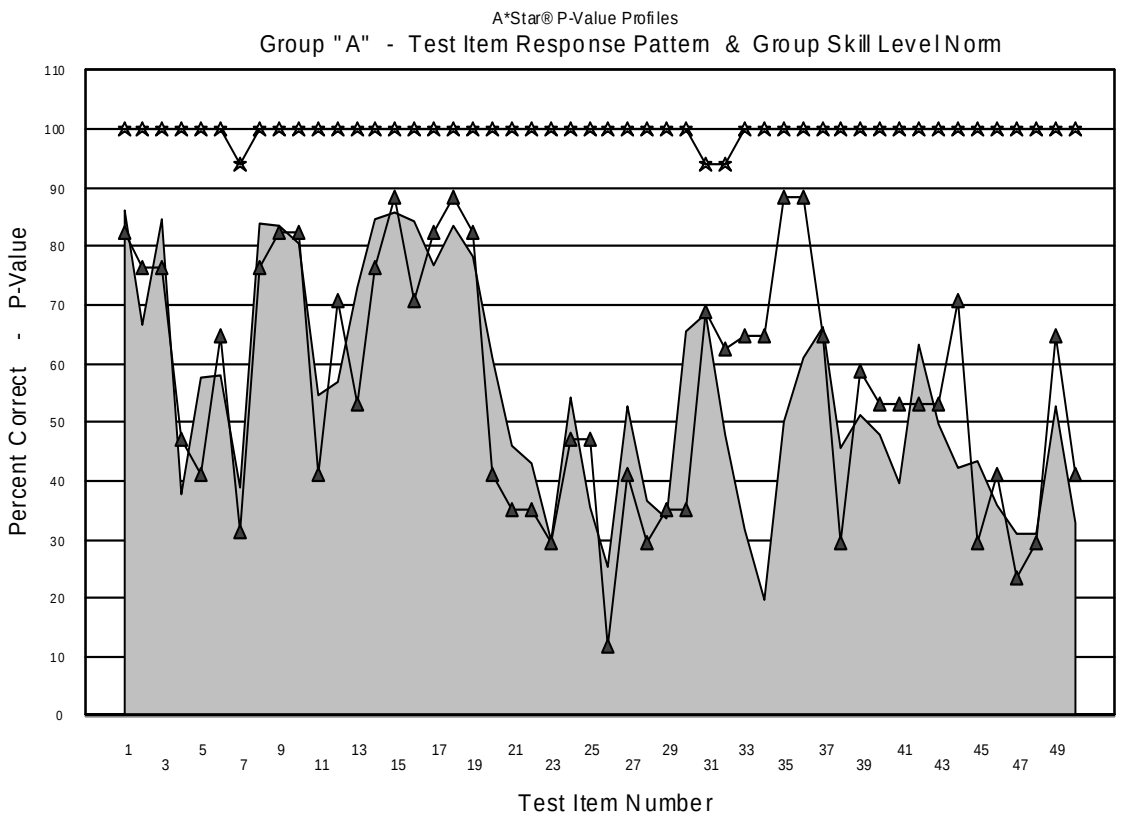
4. Identify easier test items or sections to be worked on and more difficult items to be guessed at?

5. Provide guessing 'strategies' – i.e., choose a middle size answer on math tests?

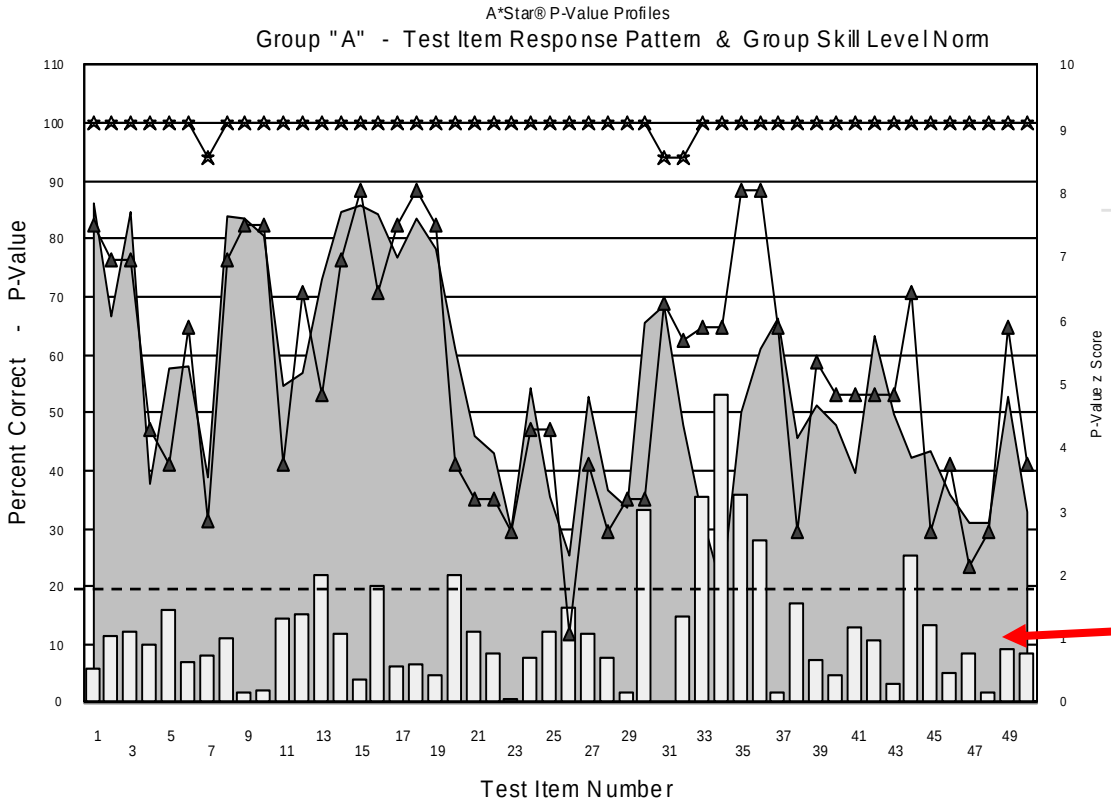
6. Give test takers extra time at the end of the testing session to fill in any remaining answer blanks?

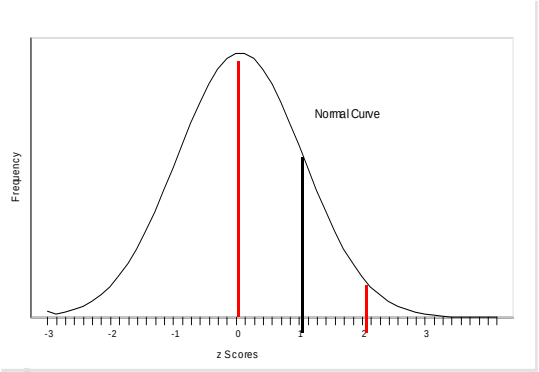
7. Suggest, encourage or insist on guessing?

Group “A” - Irregular Group Response Pattern



Group A - Exceptional Item Performance

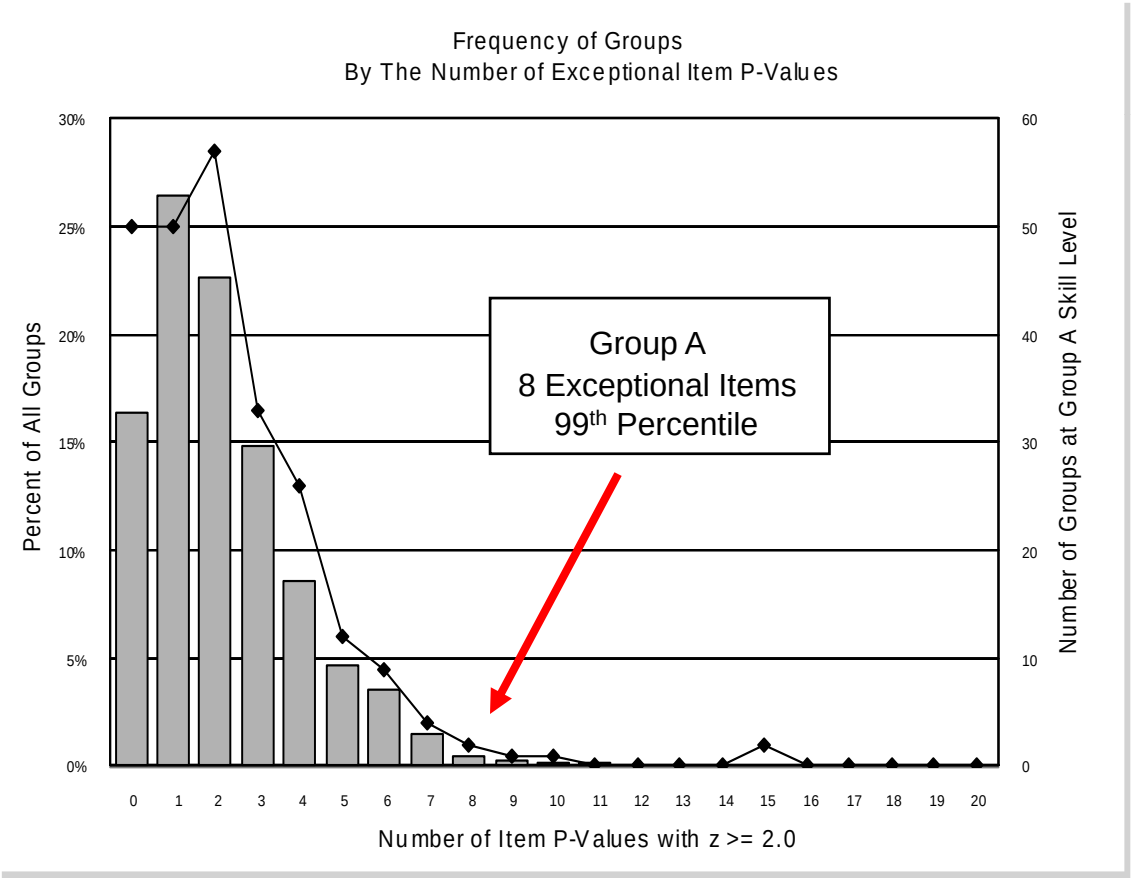




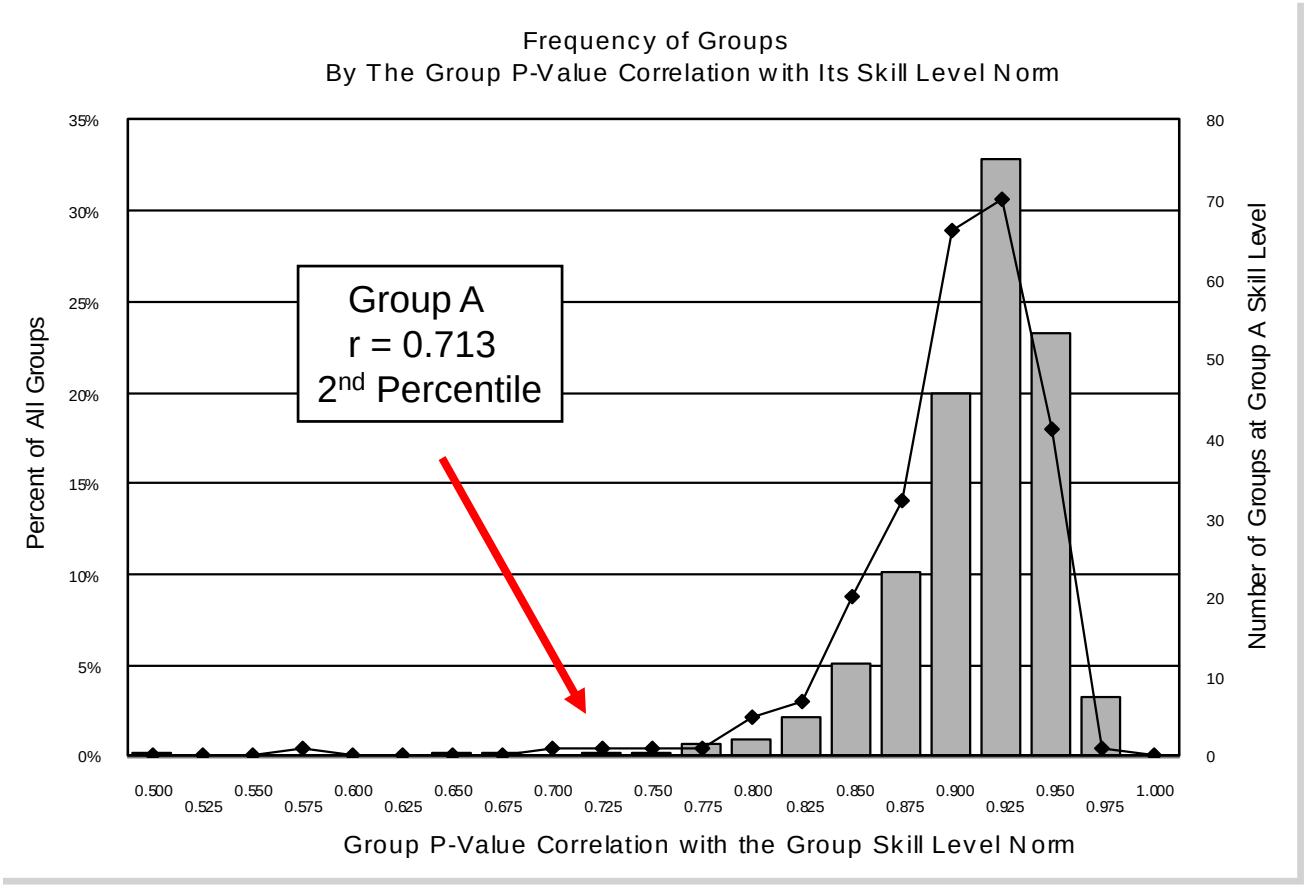
2 z Scores

Test Item
P-Value
z Scores

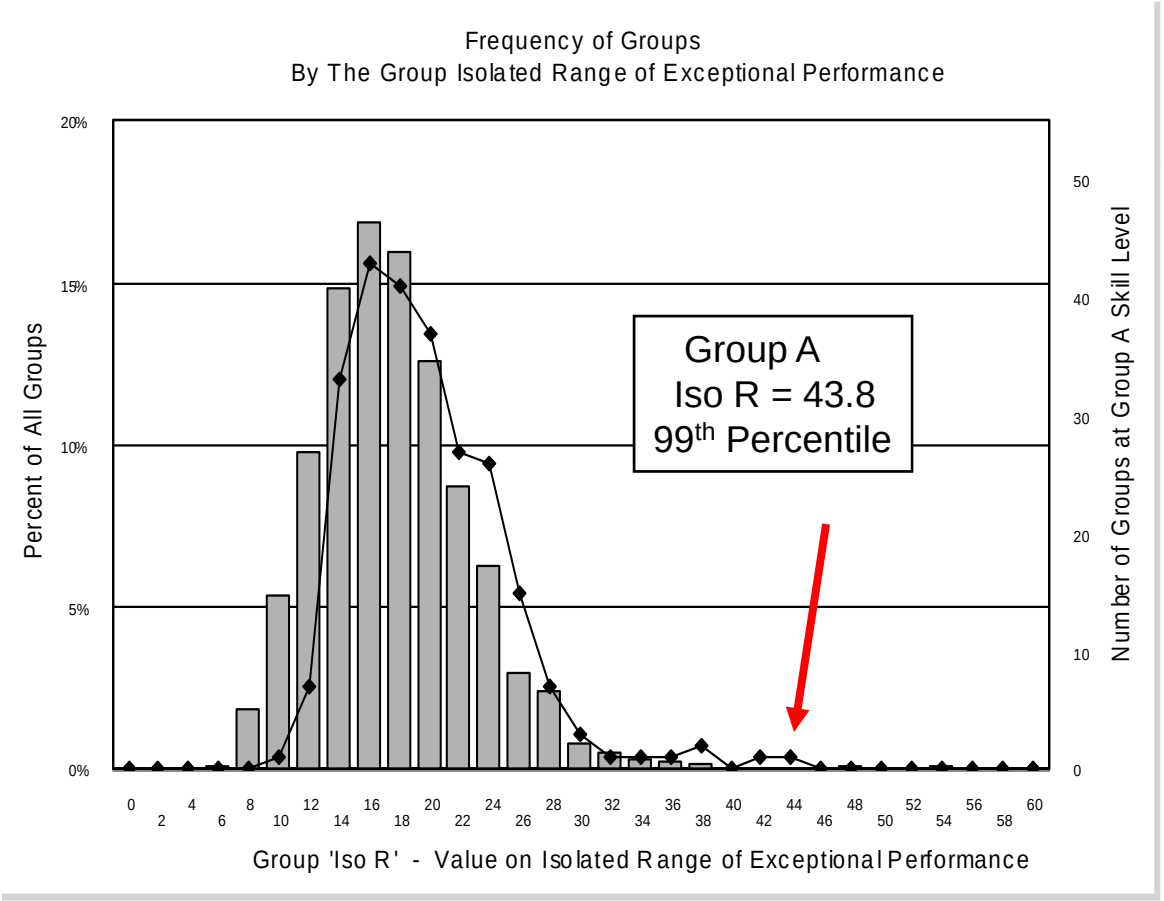
Group Frequency by The Number of Exceptional Item P-Values



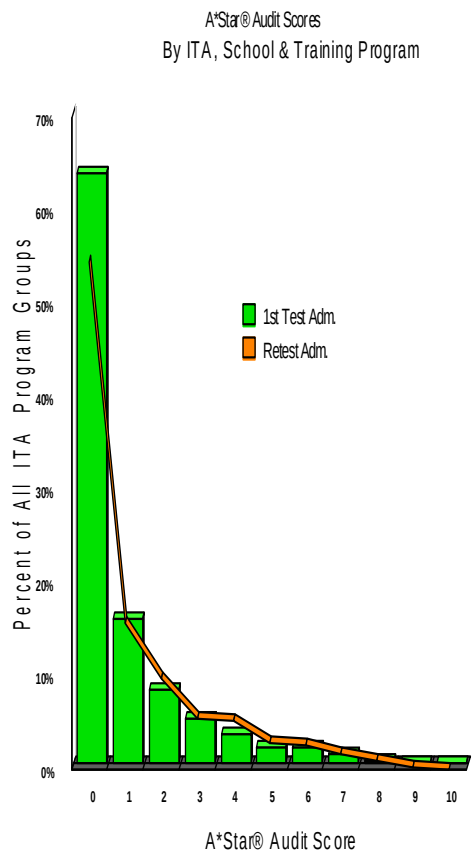
Group Frequency by The Group P-Value Correlation With The Group Skill Level Norm



Group Frequency by Isolated Range of Exceptional Performance



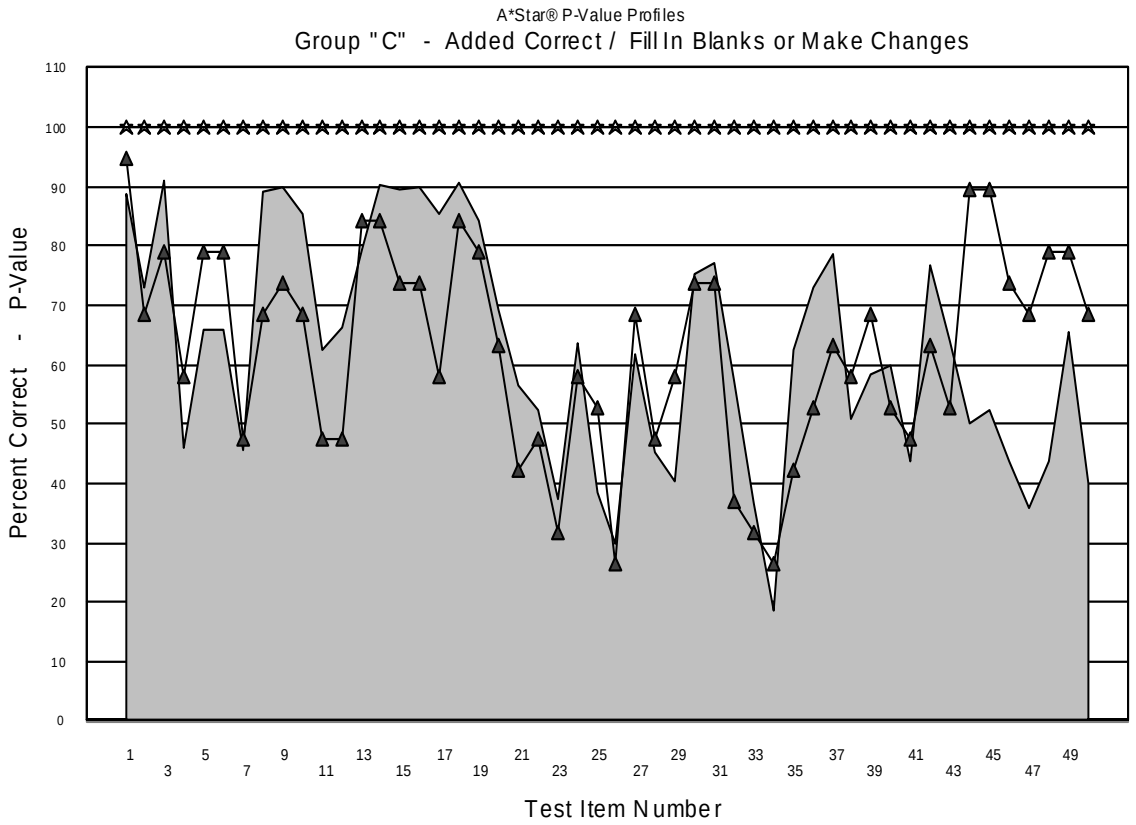
A*Star® Audit Scores



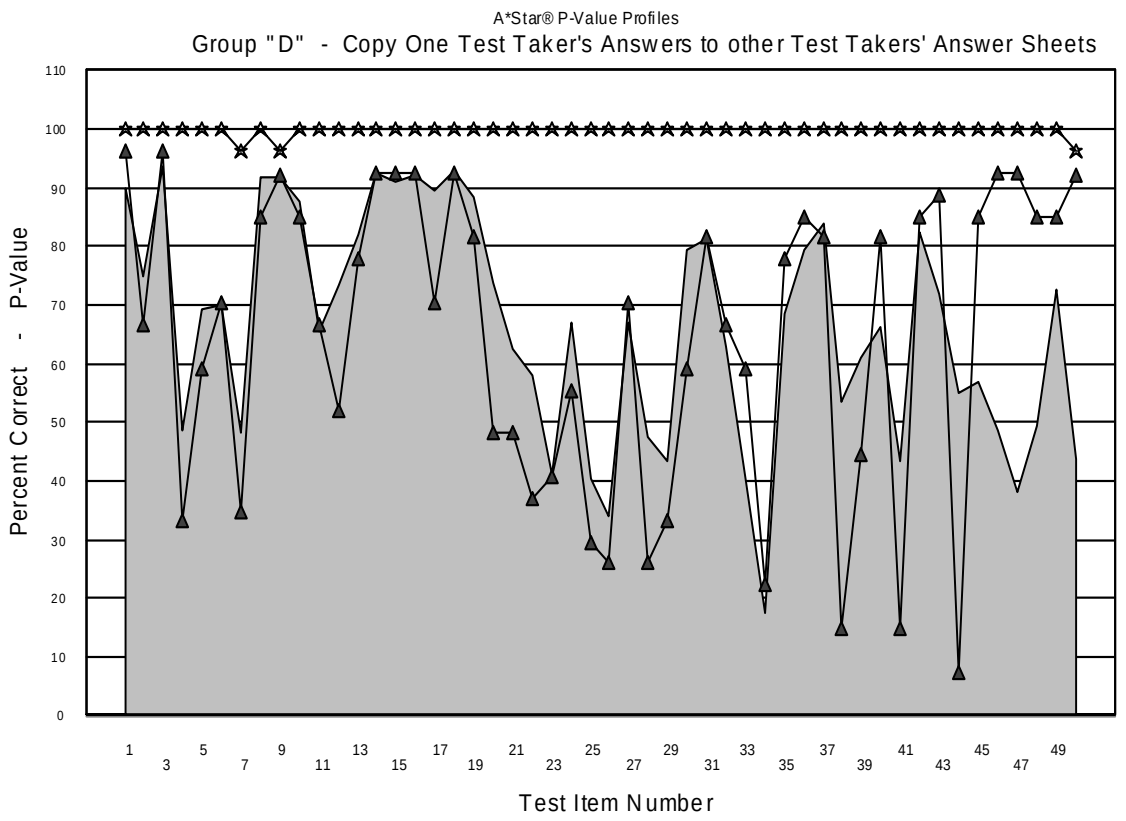
Weighted sum of multiple measures of response pattern consistency, including:

- P-Value Correlation with Skill Level Norm
- Balance of Early and Late Test Performance
- Isolated Range of Exceptional Performance
- Exceptional Item Performance
- Extended Time
- End of Test Participation

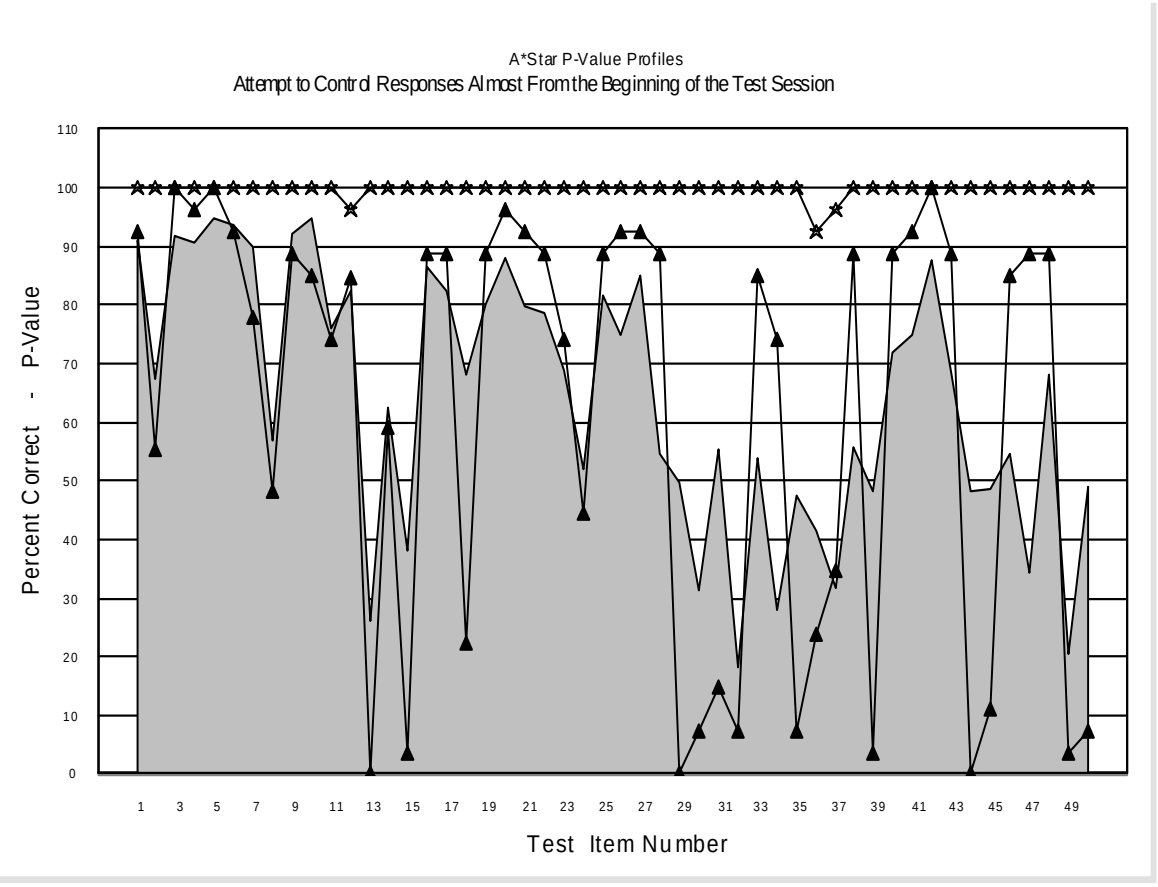
Proctor Adds / Changes Answers at the End of Test



Proctor Copies Answers From One Answer Sheet to Many Others

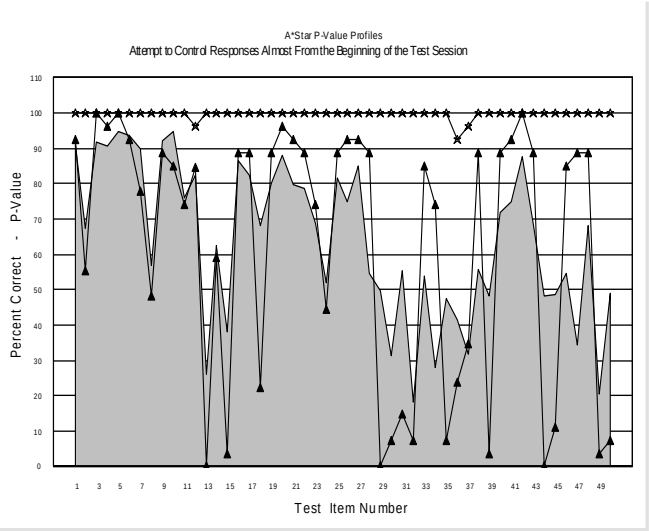


Attempt to Control Responses Over the Majority of the Test Items

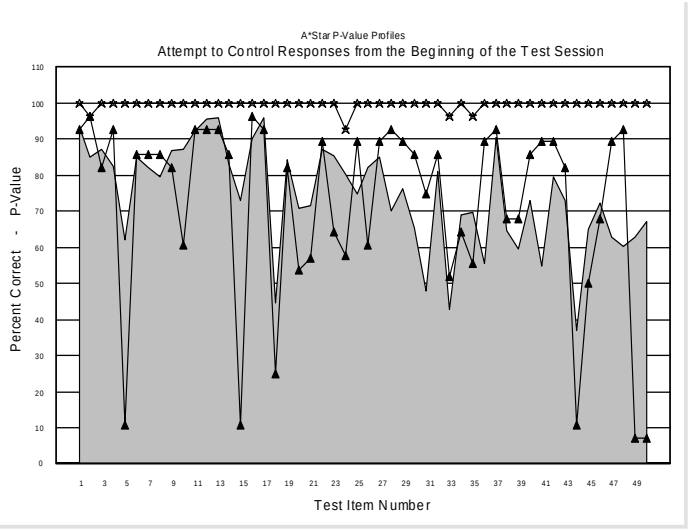


Same Proctor
Successive Test Administrations

Reading
April, 2000
Pval. r = 0.750



Math
May, 2000
Pval. r = 0.527



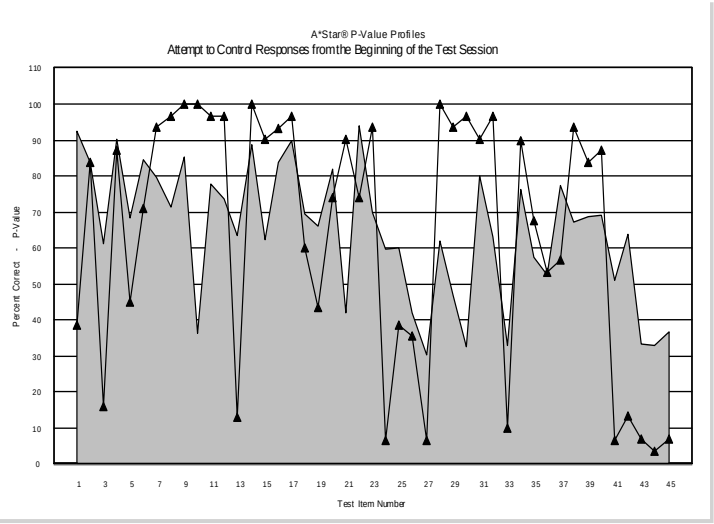
Following Year - the Same Proctor
Successive Test Administrations

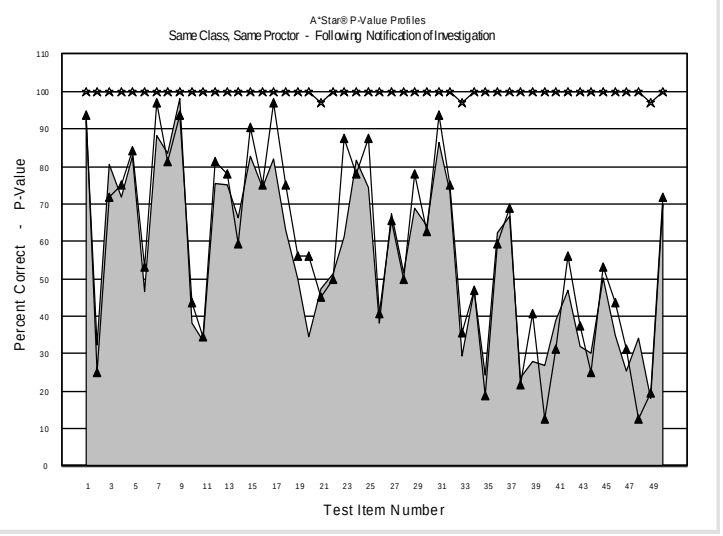
Reading
April, 2001
Pval. r = 0.487

Math
May, 2001
Pval. r = 0.947

Prior to Notification
Of Investigation

Three Weeks Following
Notification of Investigation





Subject Group Analysis

- Identify individual test takers whose test item responses underlie the Group irregular response pattern.

  Subject Group = Those most likely subject to improper influence.

- Identify the response pattern common to the Subject Group.

  Note: There may be multiple subject groups, each with different response patterns.

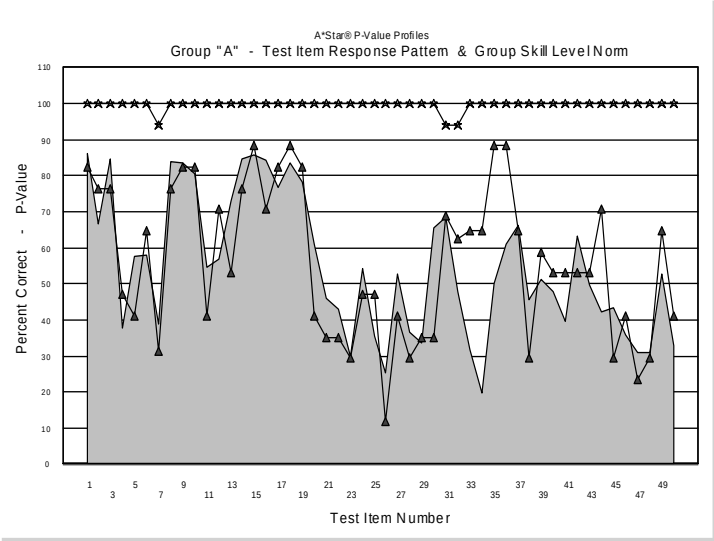
- Evaluate the probability of the Subject Group occurring in a Group of the same size and test score distribution.

Subject Group Analysis - Group A

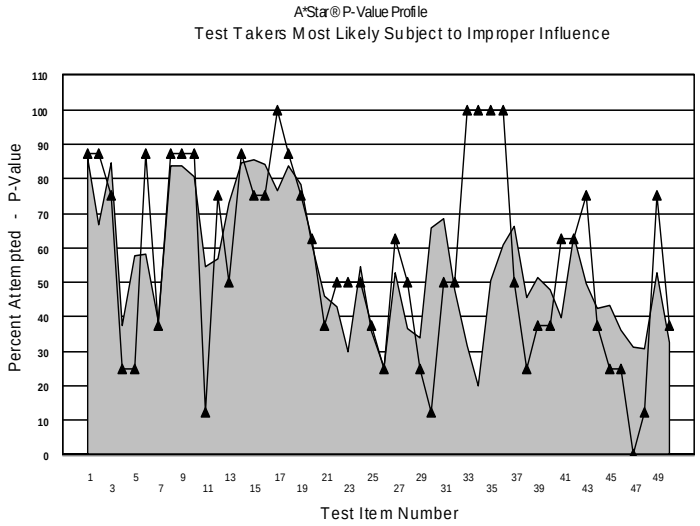
- Subject Group: 8 of 17 47.1% Probability 3.43e-9
- Expected Frequency: 0.34 of 17 2.0%

  95% Conf. Interval: 0.15% - 19.5%

Full Group n = 17



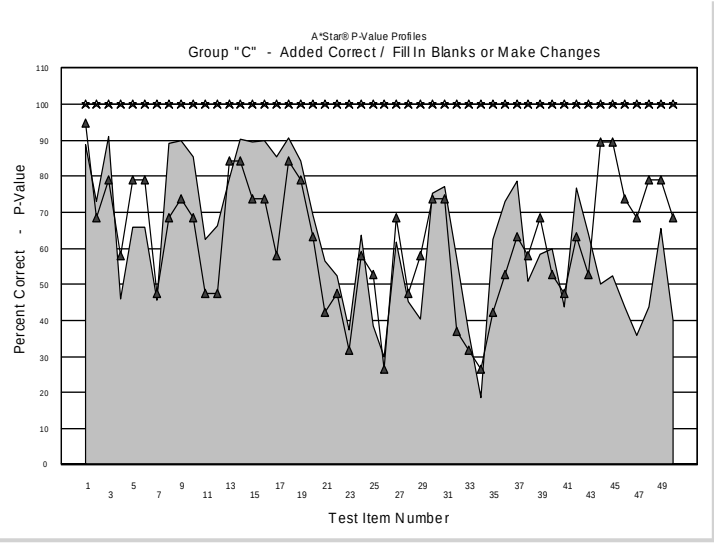
Subject Group n = 8



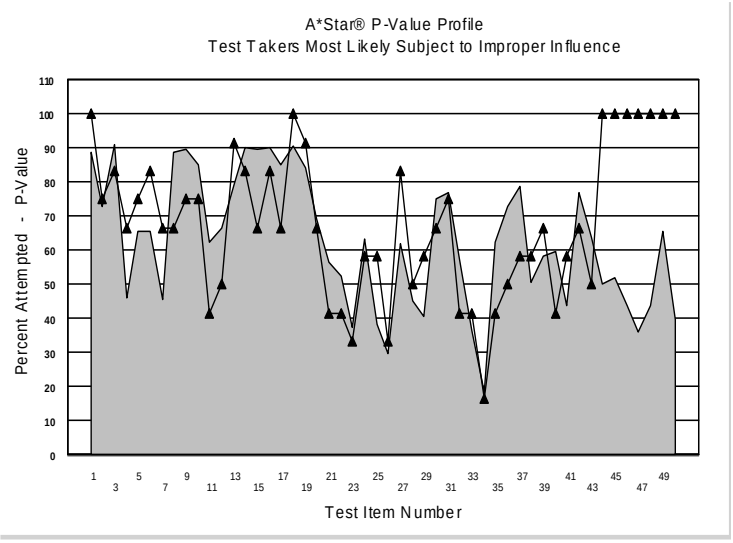
Subject Group Analysis - Group C

- Subject Group: 12 of 19 63.2% Probability 6.68e-15
- Expected Frequency: 0.36 of 19 1.9%
  95% Conf. Interval: 0.13% - 17.65%

Full Group n = 19



Subject Group n = 12

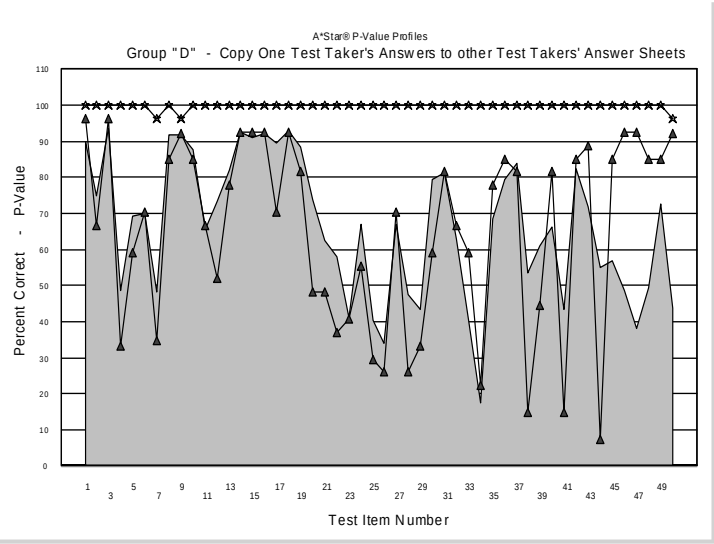


Subject Group Analysis - Group D

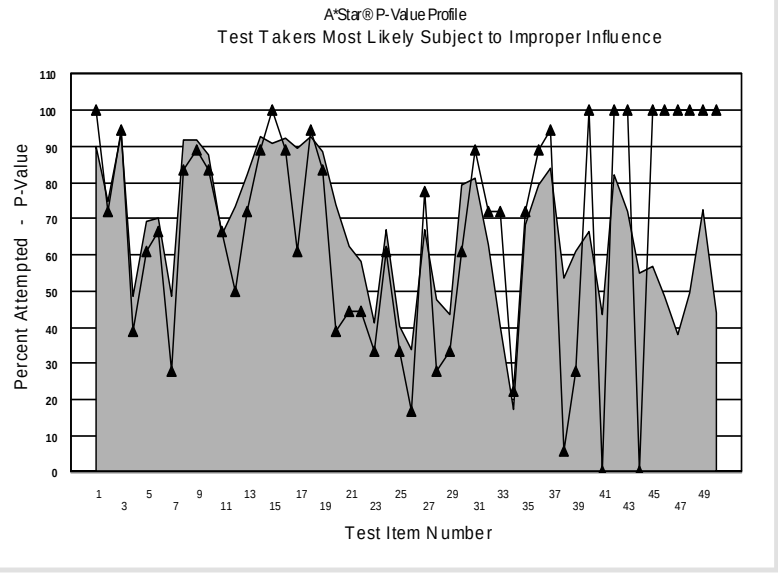
- Subject Group: 18 of 27 66.7% Probability 2.22e-16
- Expected Frequency 0.04 of 27 0.14%

  95 % Conf. Interval: 0.09% - 12.77%

Full Group n = 27



Subject Group n = 18

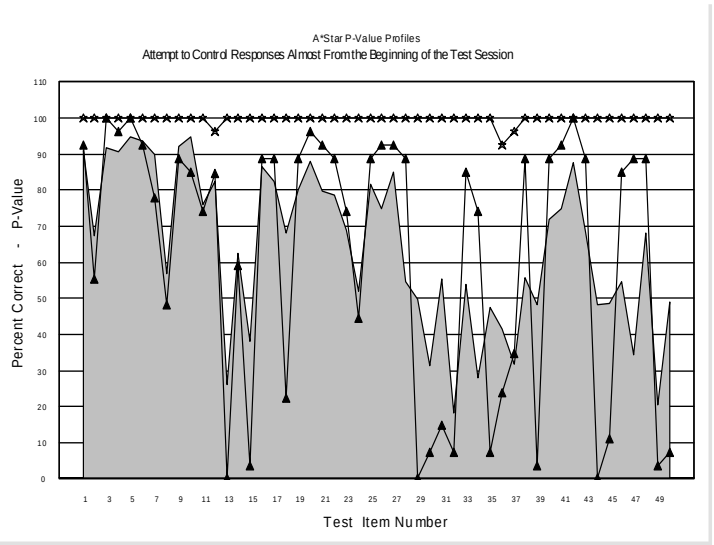


Subject Group Analysis - Group E

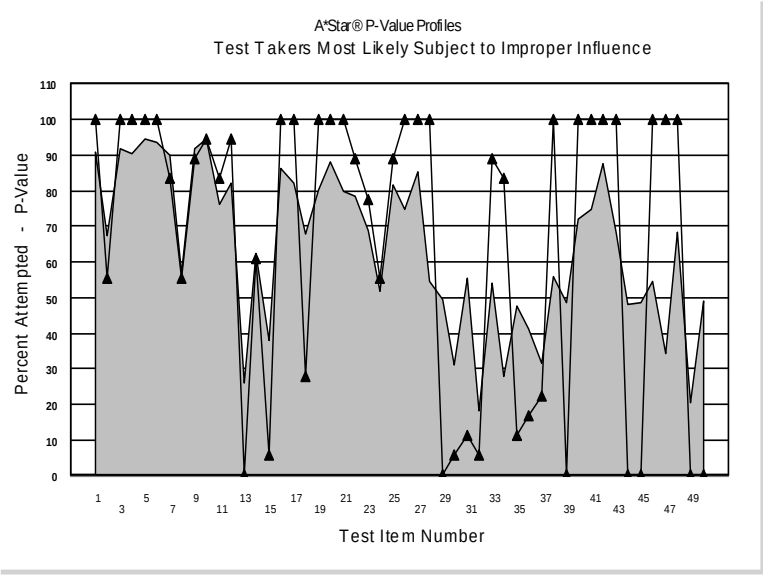
- Subject Group: 20 of 27 74.1% Probability 4.99e-80
- Expected Frequency: 0.0009 of 27 0.00%

  95% Conf. Interval: 0.00% - 12.77%

Full Group n = 27



Subject Group n = 20



How Do We Know It's Cheating?

- High level of statistical improbability for irregular response patterns.
- Consistency between the nature of the irregular response patterns and the proctor actions reported by investigators.
- Proctors identified as having improperly administered tests are up to 20 times more likely to be identified again.
- Specific irregular response patterns tend to be unique to specific proctors, often repeating for the same proctor across different administrations of the same test.
- Patterns of improper influence are common across the settings in public schools & vocational schools, but are either wholly absent or rare in industry job applicant testing.

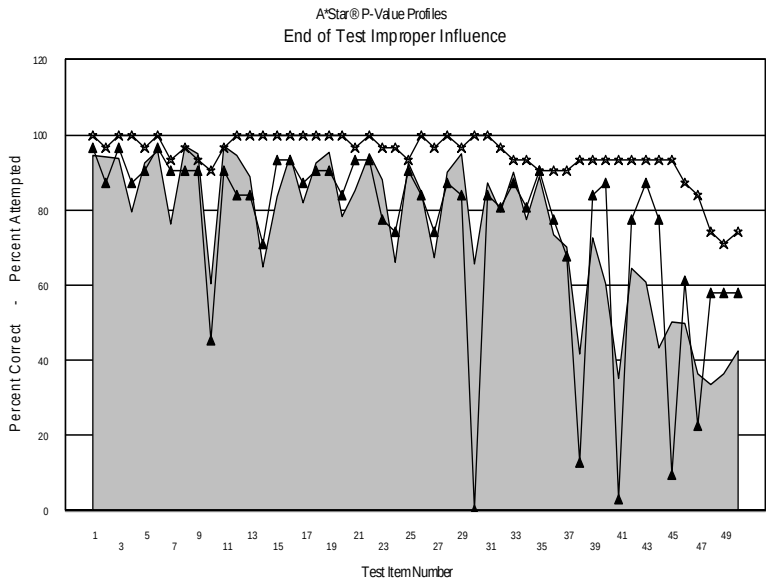
Pattern Repeated Across Multiple Administrations

This ATB test Group has been accumulated from 21 separate test administrations conducted over 19 months

90.3% of test takers choose the same incorrect alternative to Item 34.

Only the proctor is common to all test administrations.

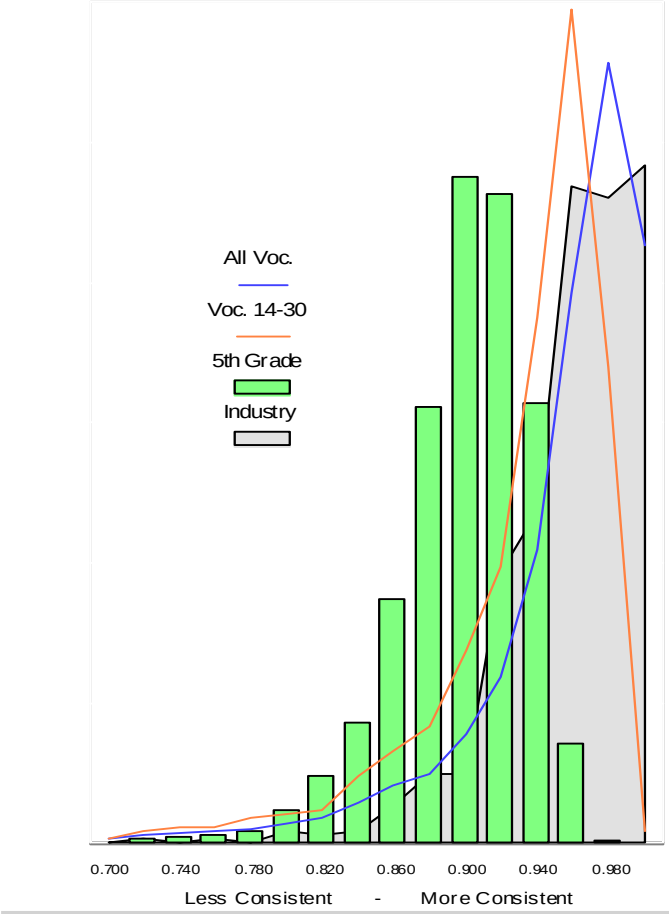
ATB Group n = 31



Distribution of P-value Correlations with Skill Level Norm

Tests: Reading - Verbal Skills

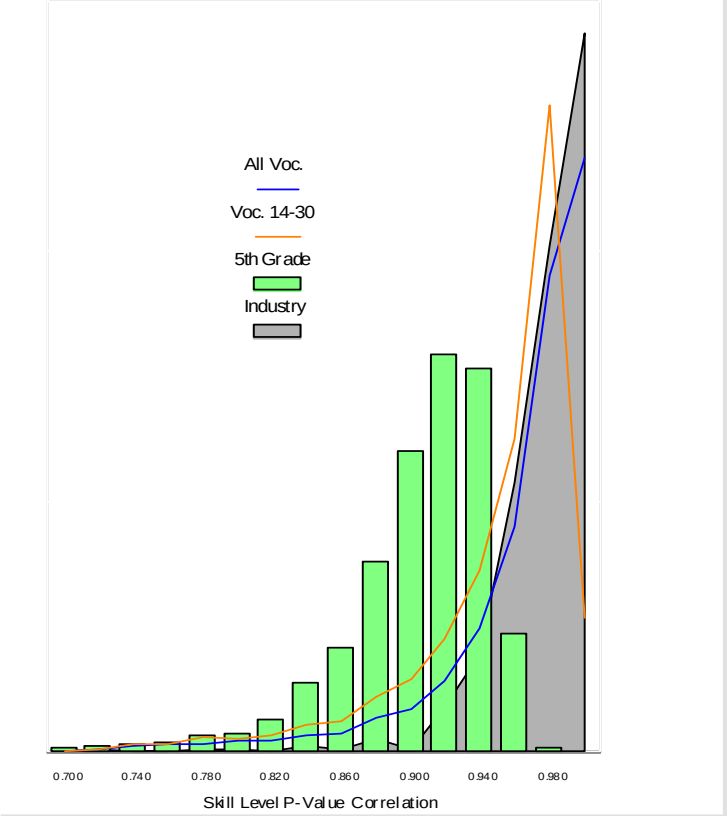
- Industry (n = 691)
  - Median 0.950
  - 25th Percentile 0.934
  - Below 0.800 0.9%
- Vocational Schools (ATB, n = 4,982)
  - Median 0.940
  - 25th Percentile 0.930
  - Below 0.800 3.9%
- Public Schools (Grade 5, n = 2,476)
  - Median 0.888
  - 25th Percentile 0.836
  - Below 0.800 2.6%



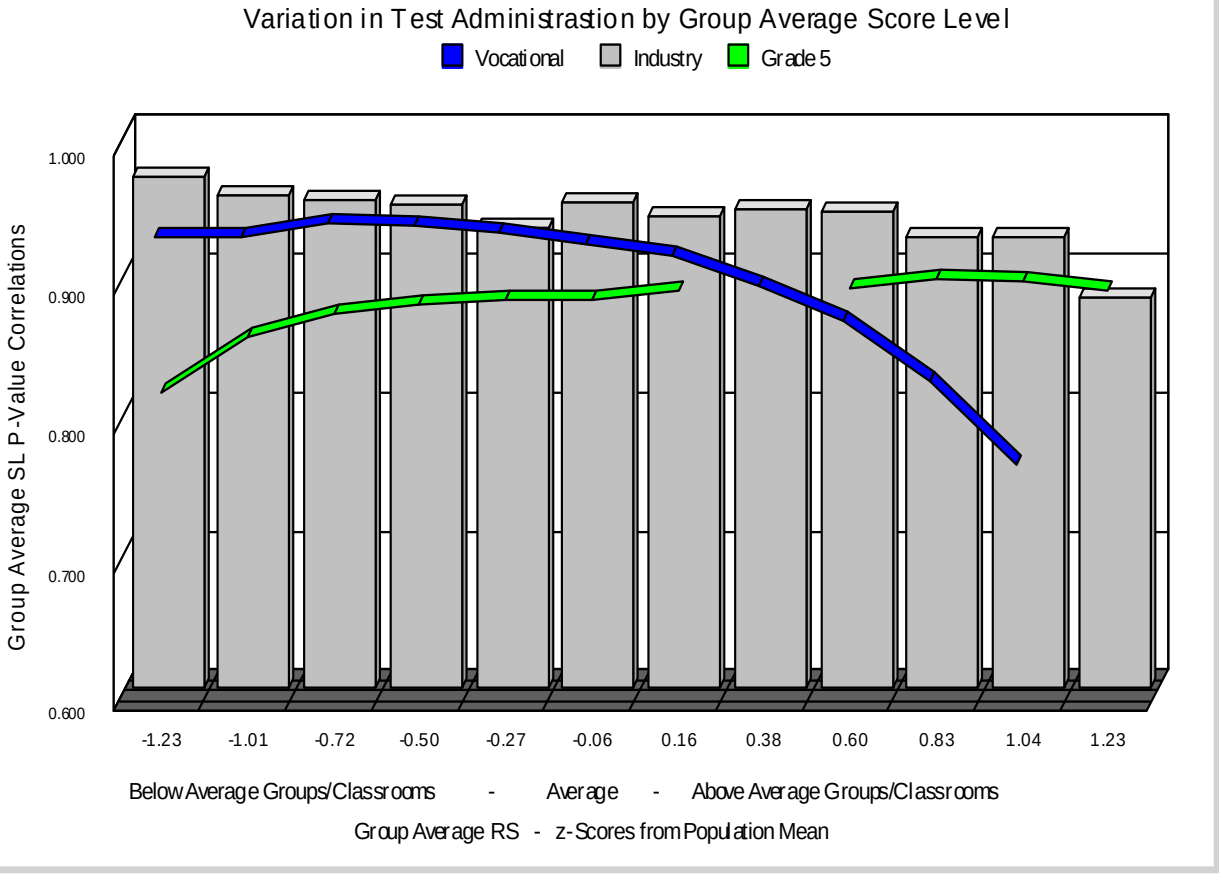
Distribution of P-Value Correlations with Skill Level Norm

Tests: Math - Quantitative Skills

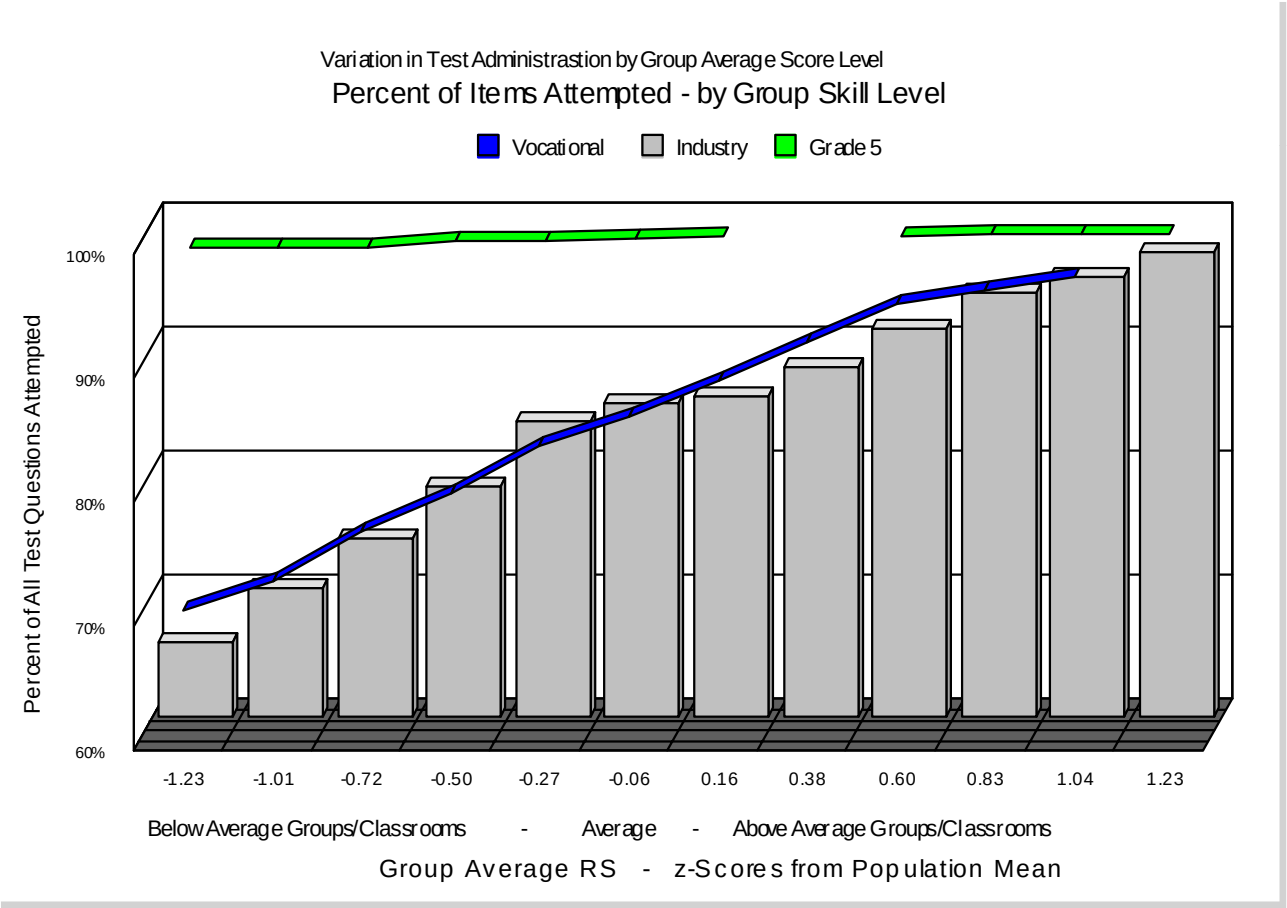
- Industry (n = 632)
  - Median 0.969
  - 25th Percentile 0.959
  - Below 0.800 0.5%
- Vocational Schools (ATB, n = 4,975)
  - Median 0.952
  - 25th Percentile 0.944
  - Below 0.800 3.3%
- Public Schools (Grade 5, n = 2,624)
  - Median 0.894
  - 25th Percentile 0.855
  - Below 0.800 4.0%



Average Group P-value Correlations By Skill Level & Assessment Setting



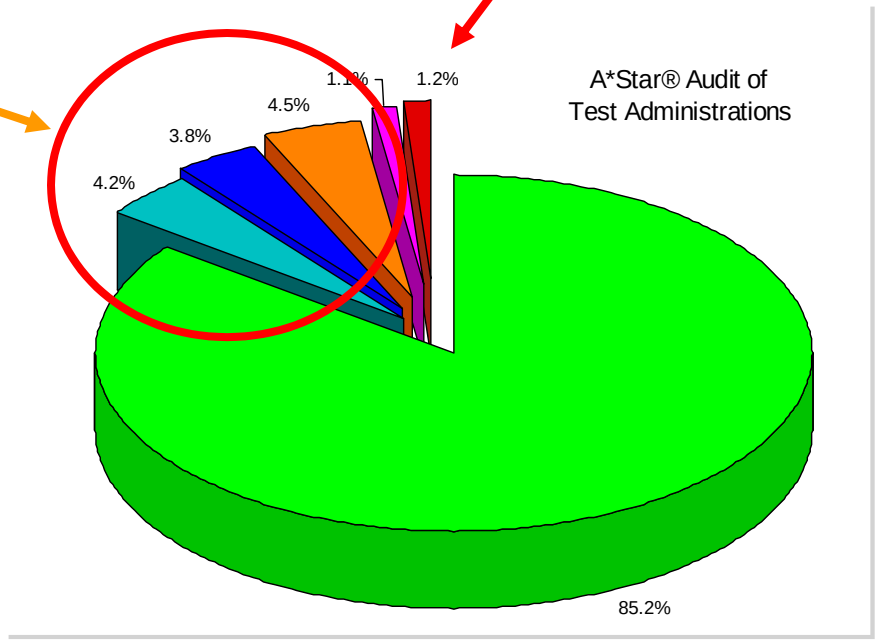
Percent Attempted
By Skill Level & Assessment Setting



Rogue Test Administrators - Extreme Examples From a Range of Improper Test Administrations

Less Severe,
But more Numerous
Problems

Rogues



Thank you for attending!

Please remember to complete the session evaluation before leaving