A*Star® Audit Review

Description of the A*Star Audit Method

Example Audit Report 4.08
and
Class Response Pattern Profiles

December 2008

A*Star Audits, LLC - 366 Clinton Street, Brooklyn, NY 11231-3702
Tel: 1-718-624-4216 - Fax: 1-718-624-1354 - email: eliotlong@astaraudits.com

Description of the A*Star Audit

Table of Contents

Page

Misadministration in standardized group testing 3

The A*Star Audit

A*Star response pattern norms 4-5
Comparison of class performance to the skill level norm 5-7
Recognizing teacher/proctor influence 7

A*Star Measures of Response Patterns

A*Star Audit score 7-8
Early-late performance 8
Correlation of class and skill level p-values 8-9
Exceptional item performance 9
Isolated range of exceptional performance 9
Performance over the last 5 items 10
Notes 11

Graphical Analysis - P-value Profiles

School & classes listed in the A*Star Example Audit Report 4.08
School 102 12-13
School 109 14-15
School 337 16-17
School 435 18-20
School 436 21-24

The A*Star Audit applies a unique method of analysis to collected test item responses to evaluate proctor compliance with standardized test administration. The following pages provide a description of the A*Star method and a set of 19 graphical analyses of the relationship between individual class test item response patterns and their appropriate skill level (peer group) norms. These graphical analyses correspond to the classes listed in the A*Star Example Audit Report 4.08. In each case, the data represent real test results from real classes and schools, but the class and school identification numbers are assigned by A*Star and have no relationship to the original.

Misadministration in Standardized Group Testing

Misadministration of tests may occur for many reasons, including a poor testing environment, misdirection of test materials, unclear test instructions, and other causes outside of the control of the teacher/proctor. Yet, the far more frequent cause found by the A*Star Audit is teacher intervention into the test work behavior of their students. Within this area, the most frequent teacher intervention is unplanned, a reaction to the struggle of their students with the test material, to particular test questions and, ultimately, to the students' inability to address all questions during the time limit of the test.

The most common condition, where there is teacher intervention, is for a modest level of assistance to a few students with a very few test items. This intervention is rationalized by some education writers[1] due to the perceived unfairness of the tests and their harsh consequences. Unfortunately, the pattern in some classrooms reveals that this modest level of intervention, once begun early in the test, grows, involving more students and test items as the testing progresses.

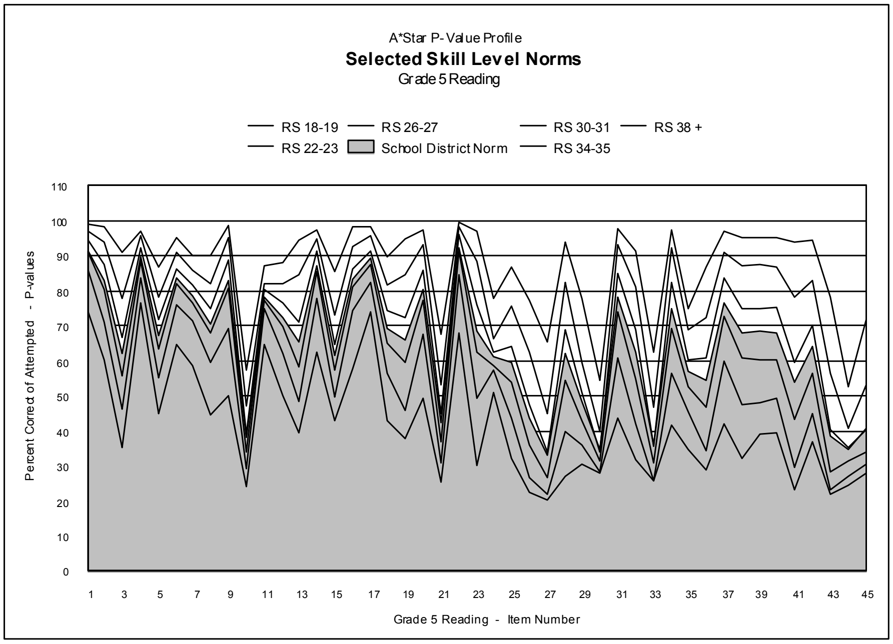
The A*Star experience indicates that very few teacher/proctors come to school with the intention of significantly increasing their students' test scores by controlling their responses. Yet, the evidence also indicates that, once teachers involve themselves in assisting their students, they are more likely to repeat the practice on subsequent testing. There is also evidence that some teachers begin with a modest level of assistance and then, over several test administrations, increase the level of their intervention. The A*Star Audit offers the opportunity to identify irregular test administrations – for whatever cause – and bring it to the attention of school administrators and teachers. This notice will encourage strict compliance with standardized test administration and proactively operate to improve future test administrations and the test score information they provide.

A*Star Response Pattern Norms

The A*Star method of review begins with a determination of group (i.e. classroom) test item response pattern norms. An example of a response pattern is the sequence of test item p-values arranged in the same order as the items are presented in the test. ‘Skill level’ norms are created based on the p-values from all groups with the same or nearly the same group average score. Separate norms are created for low performing, average performing, and high performing groups - and at many points in between. Generally, the average of the p-values for the groups at the skill level determines the norm and the variation among the group p-values determines the normal variation. Each individual group may then be compared to the norm and the comparison measured by the normal variation. This, in a nutshell, is the A*Star method granted U.S. Patent 6,960,088.

The norms based on groups of test-takers allow our analysis to recognize the dynamic of test-takers who represent a range of skill levels. Higher performing students move through the test material more quickly and successfully. Lower performing students move more slowly and less successfully, often either leaving answers blank or randomly guessing near the end of the test. This difference in performance is less pronounced at the beginning of the test session and increases as the testing progresses, changing the characteristics of the response pattern norm. When the test administration deviates from the established, standardized procedures, it is usually as a result of proctor interaction with the lower performing test-takers, changing their contribution to the group response pattern and most often appearing among the later test items. By capturing the expected dynamic of group patterns in our norms, we are better able to recognize proctor influence.

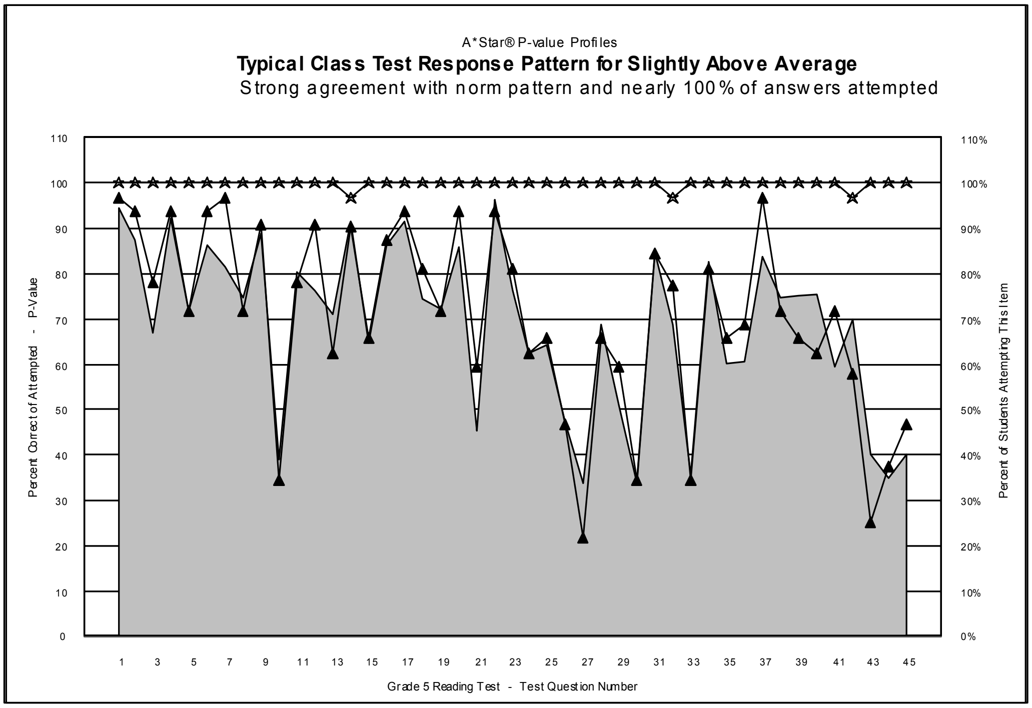
The graph below presents the response pattern norms for a range of skill levels on our example Grade 5 Reading test. The school district average is represented by the upper



margin of the gray shaded area. Norms for below average classes fall within the gray shaded area while the norms for above average classes fall above it. The pattern of the percent correct (p-value) rising and falling at the same questions for all skill levels confirms the quality of the measurement provided by the test. We notice that the norms are more closely concentrated at the beginning of the test and become more widely dispersed by the end of the test.

Comparison of Class Performance to the Skill Level Norm

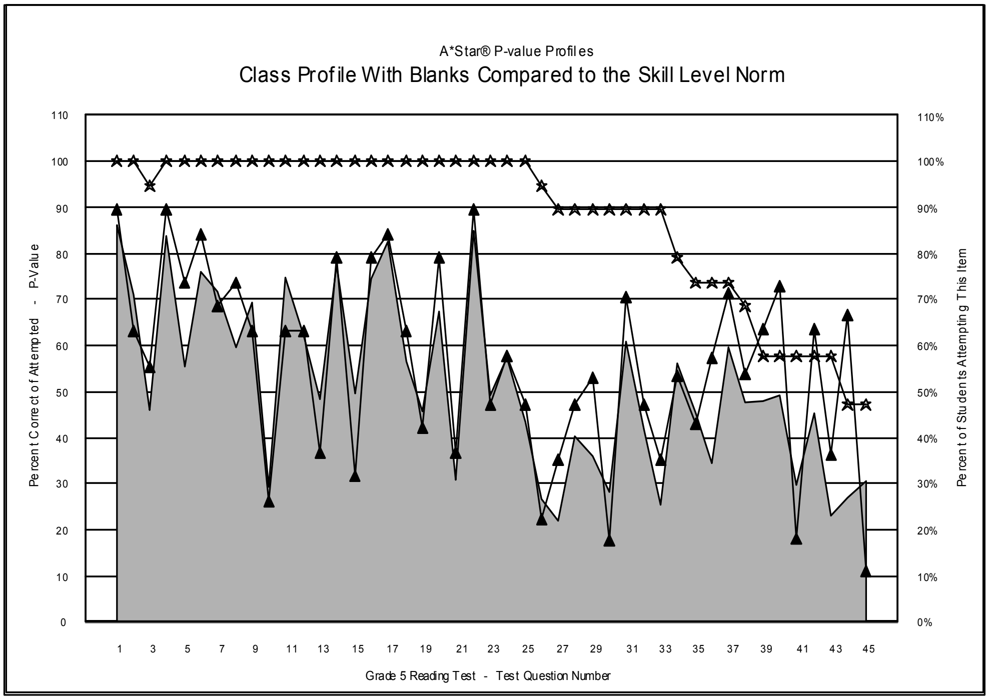
In the following graphical analyses, the skill level norm is represented by the upper margin of the gray shaded area while the class performance is represented by the solid line with arrowhead markers at each item. The percentage of the students in the class who attempted (answered) each item is represented by the solid line with star markers at each item. Sometimes this ‘attempted’ line will dip and return, indicating that either one or two students skipped the item or made an incomplete erasure, invalidating the response.


A*Star® P-value Profiles

Typical Class Test Response Pattern for Slightly Above Average

Strong agreement with norm pattern and nearly 100% of answers attempted

On relatively rare occasions (5% to 10%), the attempted line falls steeply over the later test items. This descending line reflects the volume of students who fail to complete the test. In A*Star’s experience, nearly all school districts informally encourage students to guess, if necessary, to complete an answer for all test items. A number of educational assessment writers support this practice.[2] Nevertheless, some teacher/proctors are less familiar with the practice (or object to it) and do not persist in urging their students to guess. In these classes, only the higher achieving students complete the test.



The practice of teacher encouraged guessing is a major element driving irregularities in classroom response patterns, particularly so among lower performing classes. When we compare response patterns on teacher administered tests to those from the NAEP and to those from job applicant testing by employers, we find that teacher/proctor encouragement to guess may comprise 50% or more of all test answers in low performing test-taker groups.[3] The methods used by teachers to achieve full test completion vary from classroom to classroom, sometime causing considerable confusion and lack of work effort, driving down test scores (see, for example, School 436 Class 106). The volume of test answers subject to teacher encouragement and the pressure on teachers to achieve higher test scores places a severe temptation on teachers to assist their test-takers. Strategies for improving the results from guessing (i.e. pick a middle size answer on math tests, pick one of the two most similar answers on language skills tests[4]) evolve into hints and, ultimately, direct assistance. The varying methods of teachers to apply their encouragement to guess result in a generally lower and widely varying classroom to classroom test score reliability.

Guessed correct answers also create a test score modulator: when achievement improves, guessing is reduced; gains in the correct answers due to higher skills are partially offset by reductions in the correct answers due to guessing. Similarly, a drop in skills brings an increase in guessing; the loss in correct answers based on skills is partially made up by an increase in the correct answers based on guessing. For low performing students, these offsets may be as much as 40% or more. This combination of reduced reliability and modulated scores substantially undermines efforts at program evaluation and accountability.

Recognizing teacher/proctor influence

What constitutes a significant irregularity in a class test item response pattern? The A*Star Audit applies several different algorithms to measure the consistency of the class and skill level norm patterns. These different algorithms, described below, are sensitive to different forms and severity of misadministration. No one measurement determines the condition of the administration, but rather, when considered in combination, the measurements indicate the likely character of the proctor activities and intent that created the irregularity. Norms and normal variation on each measure, based on all groups included in the Audit, provide a guide for identifying questionable and extreme values.

Improper proctor influence may be characterized as either a failure to maintain the established standardized test administration procedures or a purposeful assistance to raise student scores. While both are damaging to the integrity of the test scores and the second implies the first, the proctor's intent is revealed in the nature of the irregularity of the results. The normal pattern of responses may be broken by a randomized pattern that is inconsistent with the variation in test item difficulty, but nevertheless fails to increase the number of correct responses. This condition suggests confusion, misdirection, rushed work, and excessive guessing, but not proctor assistance with test answers. Conversely, the normal pattern may be broken by exceptional performance on specific test items or over one or more test sections. These items or areas of exceptional performance are identified as exceptional because of their inconsistency with the class performance on the remaining test items. The level of this inconsistency is a measure of the effort, and therefore the intent, necessary to create it.

Students' work on their standardized tests must be independent and self-directed to support the interpretation of their test scores. As an illustration, we liken this independence to students' play during recess on the school playground. The normal view of the playground is a cacophony of activity, with students playing either individually or in small groups of 2 or 3. If we were to see 8 or 10 students come marching across the playground in perfect step, now turning left, now turning right, always in step with each other, we would recognize the effort of an organizer. We would feel confident in the assumption that these children were submitting to the direction of someone else. So it is with test item responses.

The A*Star Audit Score

The A*Star experience in reviewing test-taker groups from many settings suggests certain fixed values on the measures of response patterns, beyond which significant proctor intervention into test-taker work behavior becomes indicated. The combination of these values are captured by the A*Star Audit Score and summarized in a score range of from 0 to 3 where: 0 = within normal variation, 1 = modest irregularity, 2 = significant irregularity, and 3 = severe irregularity. Classrooms with an Audit Score of 1 fall outside of the expected variation, but the deviation is modest (involving relatively few items and few students) and the cause may well not be due to the proctor's activities. Audit Scores of 2 suggest a significant role for the proctor in causing the irregularity, yet the impact is limited to either a relatively few test items or a few students, suggesting a lack of intent to make a major impact on the overall class results. Audit Scores of 3 suggest a dominant role for the proctor, including a purposeful effort to influence the overall class results. In the actual, full set of classroom results represented by the Example Report, the Audit Score frequencies were: 1 = 12.6%, 2 = 2.3%, and 3 = 0.6%.

The A*Star Audit score is based on combinations of the different measures of response pattern consistency with the skill level norm. These measures are described below.

Early – Late Test Performance

The Audit develops a model for forecasting the total test score based on the early (first 60%) test content and, again, based on the late (last 40%) test content. This break at 60% (i.e. the first 30 items on a 50 item test) is based on our experience that this is about as far as lower performing students get based on their own skills and therefore it is where proctor intervention will most likely begin. If the test has more than one timed session, the Audit will create a separate set of forecasts for each session.

A comparison of the two total score estimates provides a measure of the consistency of the class performance over the entire test content. When, for example, there is an effort to improve the test performance as the testing session ends, late test performance will increase disproportionately. Efforts to rush test-taker work early in the test session may lower early test performance. The significance of the proctor influence will be measured by the size of the Early-Late differential (or SPD value), but will require other measures to interpret the significance and intent.

A Correlation of Class and Skill Level P-values

A correlation of the class p-values with the norm p-values provides a 'whole' test measure. The norm for all classes on this Grade 5 Reading test (each measured against their own skill level norm) is a correlation of .892 with a standard deviation of .038. Two standard deviations below the norm will be .816. A low correlation indicates a sharp contradiction between the variation in test item difficulty and the variation in class performance; a significant number of test-takers must unexpectedly answer easy questions wrong and/or difficult questions right.

When we look over a large number of classroom results, these low correlation are rare (0.8%). The Example Report lists three classes with p-value correlations well below .750. Two of these (School 337 Class 501 and School 436 Class 102) are clear examples of serious, improper proctor influence on their test-takers' work behavior; one (School 436 Class 106) is a case of serious confusion, but unlikely to include purposeful efforts to raise scores. The Report also lists one class on the borderline, with a correlation based on adjusted p-values of .740 and based on unadjusted p-values of .790. This class is very likely to have experienced improper influence, yet it does not rise to our level of certainty. Each of these classes are evaluated, first based on its low correlation, and then on other measures.

Exceptional Item Performance

The Exceptional Item Performance values indicate the frequencies of class p-values 1.6 and 2.0 standard deviations (z scores) from the skill level norm (the .05 and .025 levels, respectively). It is these 'z-items' that suggest purposeful efforts to raise test performance. Elevated p-values may be applied to the number in the class to estimate the number of students involved. The lack of an elevated number of z-items for School 436 Class 106 confirms the confused character of the test administration that resulted in an exceptionally low p-value correlation. Conversely, the high frequency of z-items for the other low correlation classes confirms our interpretation of purposeful efforts.

Frequently, the proctor influence will be concentrated in a limited area of the test content. We may then 'fit' the class response pattern on the remainder of the test content to the skill level where we find the best agreement with the norm. When we again measure the elevation of the p-values at the effected items, the degree of influence often markedly increases.

Note that the measure of z-items includes both items with high p-values well above the norm and low p-values well below the norm. A common form of improper influence is to use the answers of one student to assist other students: both correct and incorrect answers are copied. The response pattern for School 436 Class 102 is a good example of this condition.

Isolated Range of Exceptional Performance

The Isolated Range of Exceptional Performance ('IsoR') identifies a succession of items with elevated p-values. The grouping of a relatively small number of items with high p-values creates a far more unexpected condition than would the same number of high p-values scattered throughout the test. Note, for example, School 109 Class 502 where the expectancy for Exceptional Item Performance is 26.2% while the expectancy for the IsoR is 0.1%. This IsoR value often arises where proctor assistance with a test section is indicated. This may be instruction ("Remember to invert and multiply to divide fractions.") or direct assistance with correct answers. Often, the IsoR section is at the end of the test where answers left blank may be filled in with correct answers.

Performance over the Last 5 Items

The variations found in the condition of group responses at the end of the test are consistently the largest variations found among classroom groups. It is also a major difference between the response patterns from teacher administered tests and the patterns from tests administered by the independent proctors of the NAEP. The primary difference is whether or not low performing test-takers complete the end-of-test items or leave the answers blank. In the great majority of classes, all students answer all test items. In a significant minority of classrooms (10% to 15%), low performing student leave these answers blank. The nature of the variation in test completion among classrooms at the same achievement level makes it clear that these differences are due to variations in proctoring.

With the first difference being whether or not the answers are filled in, the second difference is in the method used to obtain the answers. The most common proctoring method is to encourage students to work quickly early in the test, to save time for the later test items. There is evidence that this lowers low performing students' performance over the early test sections to a greater degree than it raises their performance over the later sections. (Ask about this evidence, it's a neat story about summer school.) Other methods range from encouraging random guessing, to extending the time limit, to assisting with hints and answers, to filling in blank answers after the answer documents have been turned in (sometimes with random guessing, sometimes with correct answers).

The A*Star Audit provides us with a measure of the class p-value for the end-of-test items (usually the last 5 items). The p-value is the number of correct answers over these items divided by the number of answer attempts. This p-value is compared to the same value for all classes at same the skill level and a z score is determined. The p-value, the z score and the percent of the end-of-test items attempted is reported. Note that this percent attempted is, for 5 items as an example, the number of attempts over the 5 items divided by the number of students in the class times 5 (Attp./(n x 5)).

The percent attempted is an important key to this last section. As noted, in our experience, most classes have very nearly 100% completion. Those classes that do have a significant number of questions left blank (i.e. School 436 Class 105) are not likely to have been subject to efforts to raise scores. The students in these classes who did complete the final questions are likely to be the highest achieving students in the class and therefore achieve a high percent correct than the skill level norm. We tend to ignore high z scores for classes with low test completion.

Notes:

1. Nichols, S. L. & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press, p. 34

2. Examples of assessment writers who recommend encouraged guessing are:
   Linn, R. L. & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle, NJ: Prentice-Hall, 354-355.
   Mehrens, W. A. & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (4th ed.). New York: Holt, Rinehart and Winston, 461-465
   Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill. 340-348.

   The recommendations of these and other writers are based on research conducted on samples of test-takers prior to the 1980's. This research does not consider the consequences of encouraged guessing in practice or on students substantially challenged by the difficulty of the test.

3. Long, E. R. (2008, June). *Searching for DIF, drift, and growth among the deck chairs*. Paper presented at the Conference of the Fordham Council on Applied Psychometrics, New York, NY.

4. Loulou, D. (1995). Making the A: How to study for tests. *ERIC Digests*. ERIC Document Reproduction Service No. ED385613.

   The internet provides a great many resources for suggestions on how to strategize guessing on tests. None of these are likely to be successful on professionally developed tests, but are likely to reduce test-taker work effort and lower scores.

Graphical Analysis - P-value Profiles

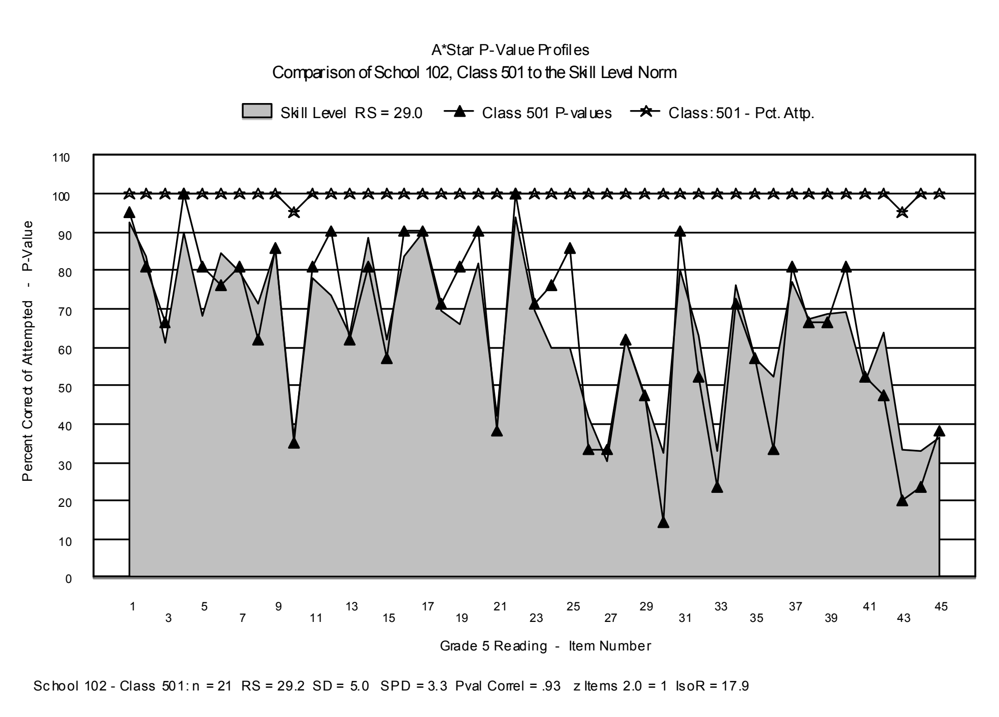
In the graphs that follow, we present examples of good, questionable, and clearly improper test administration. In each of the graphs, the gray shaded area represents the skill level norm set by all other classes at the same, or very similar, class average score. As you look through the graphs, you will see the norm rise and fall with the class average score, and yet the class will – most often – track right along with its norm.

In the graph below, for School 102 – Class 501, there are a few points where the class p-values rise distinctly above or fall distinctly below the norm. Most often, when the class p-values fall below the norm, it is at difficult items where the teacher has encouraged the students to "not waste time, guess and move on." When the p-values rise above the norm, it is most often at easy items where the students have gained time by guessing at the difficult items.

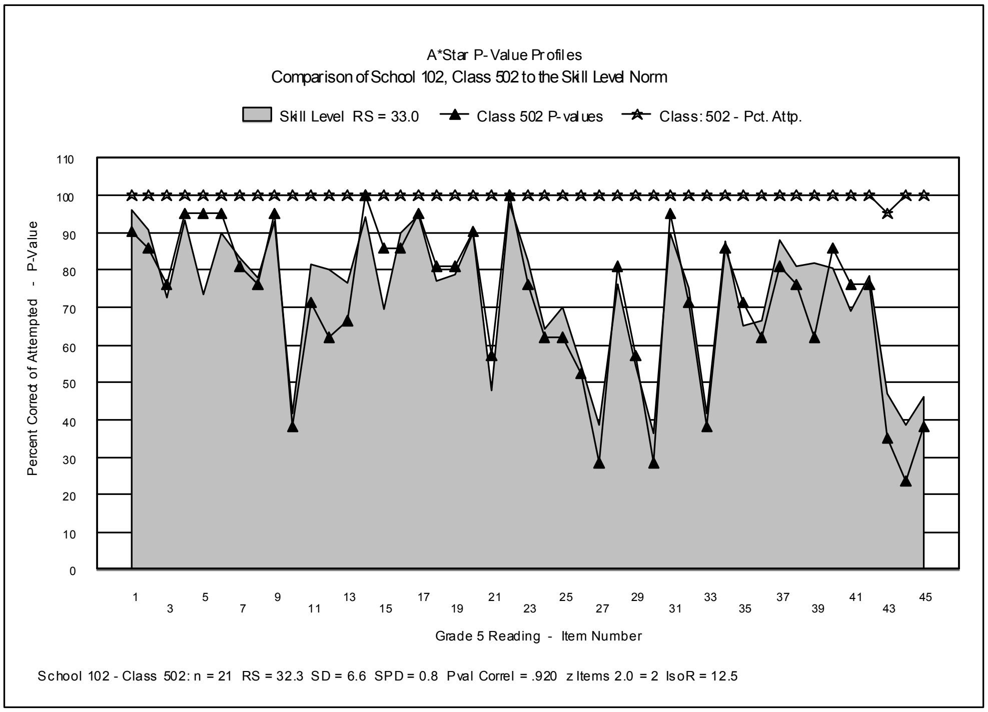
When teachers assist students, it is usually with the more difficult items. Occasionally, in very serious cases, the teacher will attempt to mask their assistance by copying answers from one of their better student's answer document. Both right and wrong answers will be copied, dramatically raising and lowering the p-values.

School 102

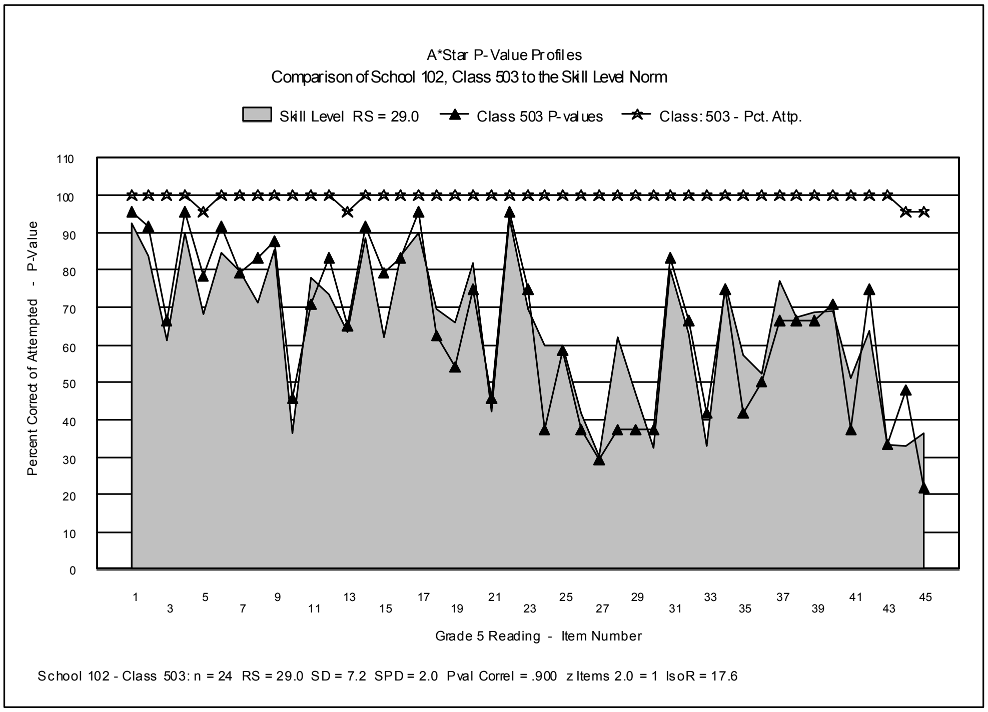
School 102 – Class 501: Good agreement with the norm (r = .93).



School 102 – Class 502: Excellent agreement with the norm for a modestly above average class.



School 102 – Class 503: Good agreement with the norm (r = .90), perhaps more careful work early and more rushed work later.

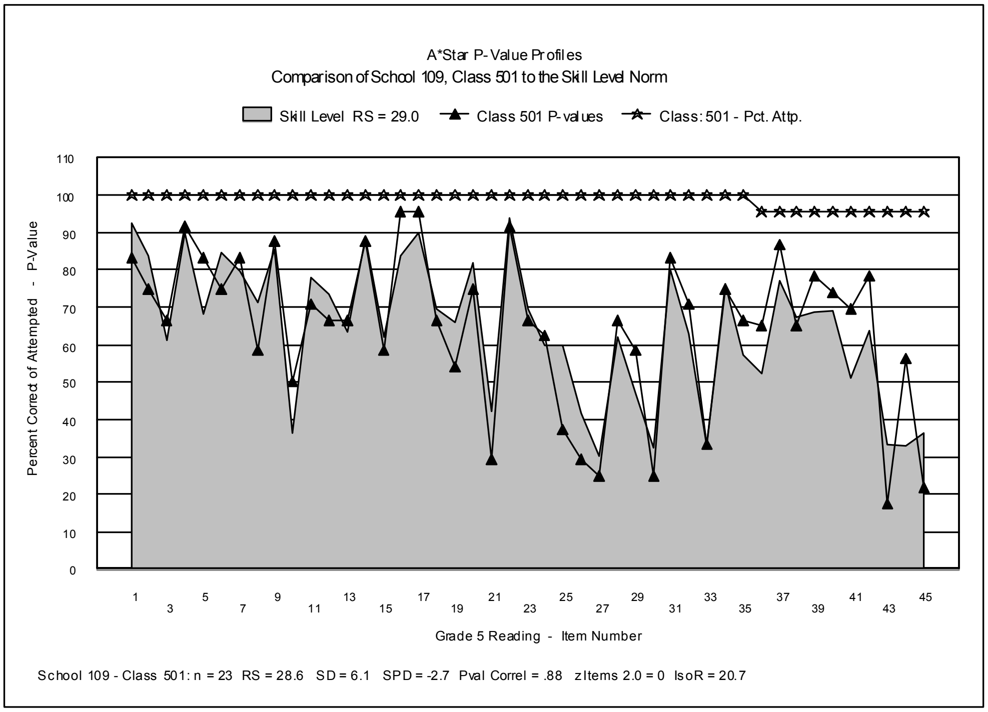


School 109

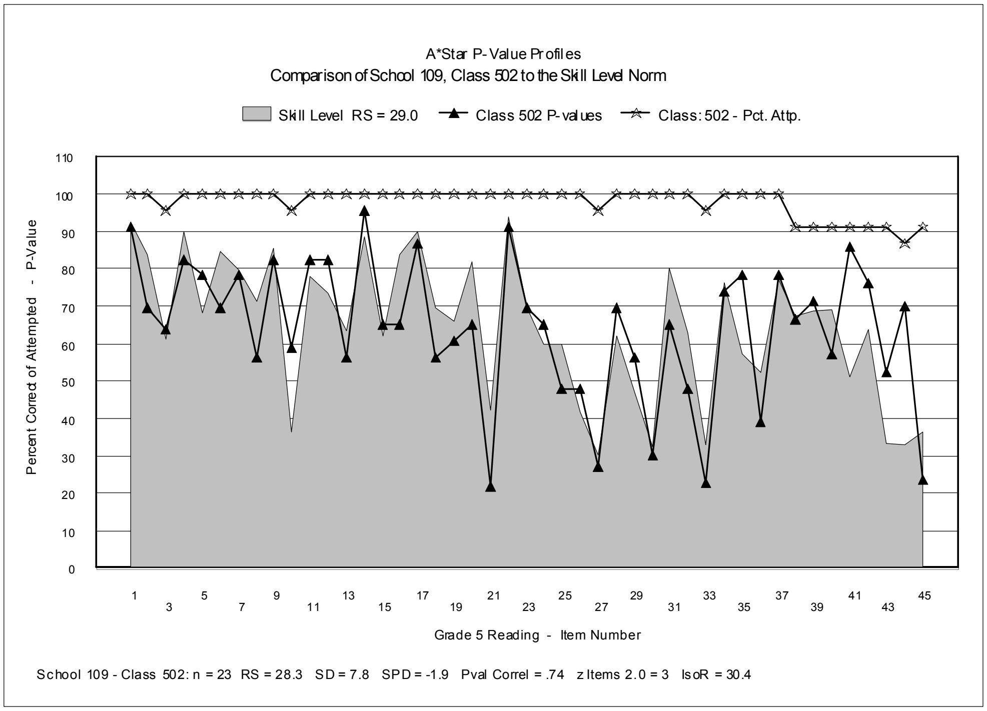
The irregularity found in the response patterns of the three classes at School 109 raises a concern due to their similarity. All three classes perform better over the end of the test than we would expect given their performance over the early test items. Generally, when the class p-value correlation with the norm is above .85 and the higher late test performance is modest, the cause may simply be either rushed work early in the test or a favorable outcome from guessing. When the late test performance is substantially higher (about 3+ points), we are concerned for the proctor either allowing extended time or providing assistance. We note, in Class 502, that 3 students correctly answered items 41-44 (the test has 45 items) although they performed at very modest levels early in the test.

On the basis of this review, it would be appropriate for the school to review, with all teachers, the procedures to handle the timing and collection of test materials at the end of the test administration.

School 109 – Class 501: Generally good agreement with the norm (r = .88), but an unlikely, elevated performance over items 35 – 41. Suggestive of extended time.

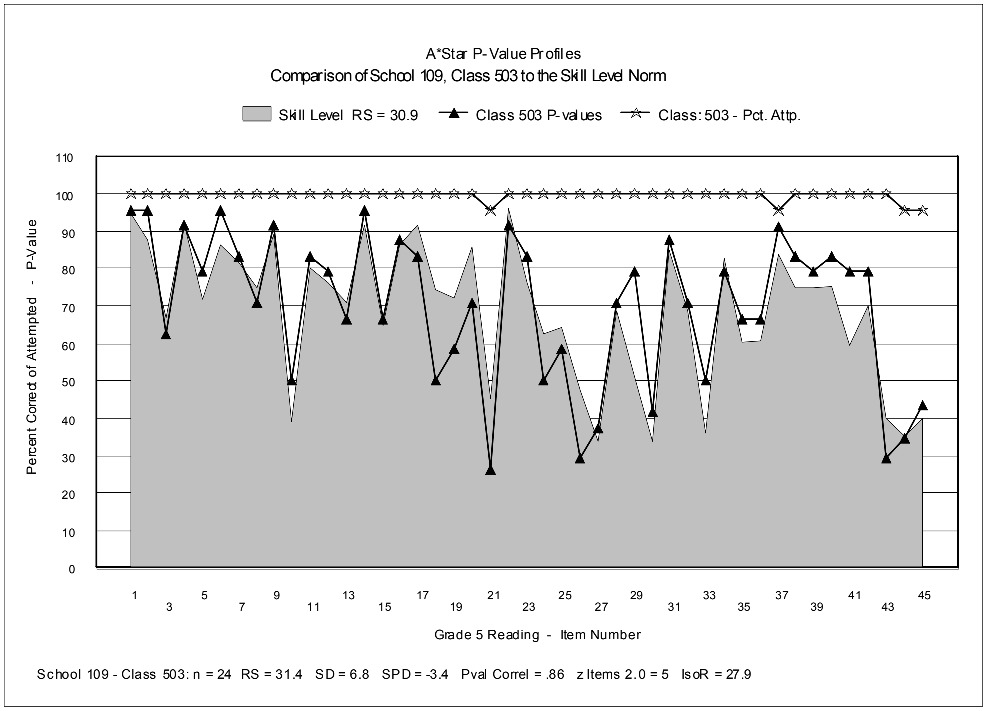


School 109 – Class 502: Poor agreement with the norm (.74) and markedly elevated p-values over items 41 – 44. This is suggestive of improper assistance to 3 to 5 students.



School 109 - Class 502: n = 23 RS = 28.3 SD = 7.8 SPD = -1.9 Pval Correl = .74 z Items 2.0 = 3 IsoR = 30.4

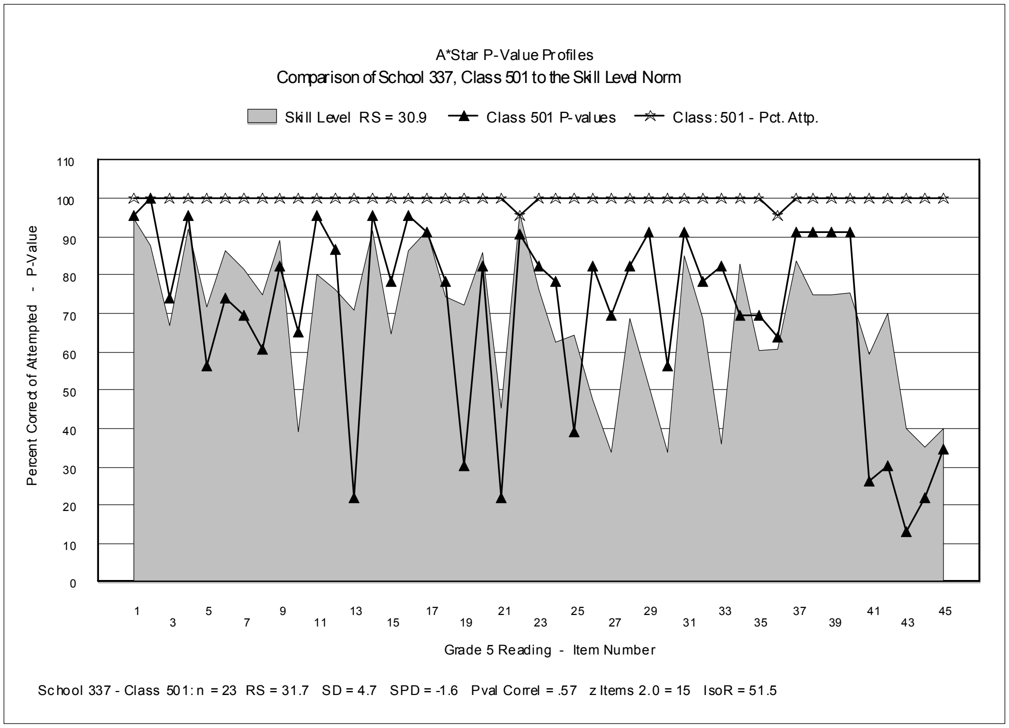
School 109 – Class 503: Fair agreement with the norm (r = .86), yet unlikely elevated performance over items 35 – 41 – the same region as with Class 501.



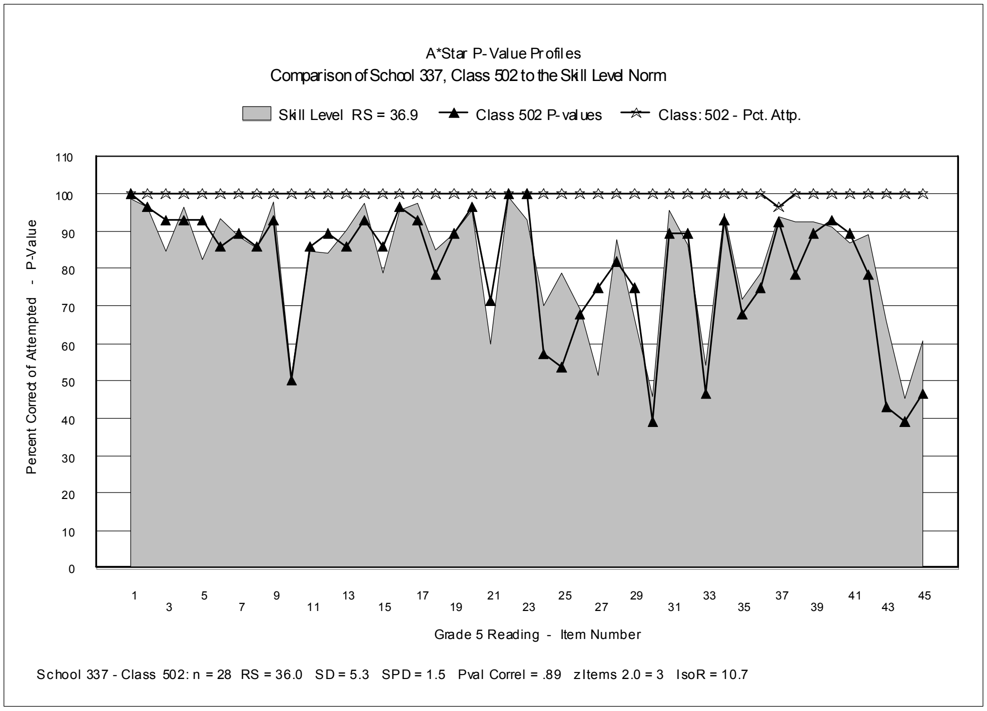
School 109 - Class 503: n = 24 RS = 31.4 SD = 6.8 SPD = -3.4 Pval Correl = .86 z Items 2.0 = 5 IsoR = 27.9

School 337

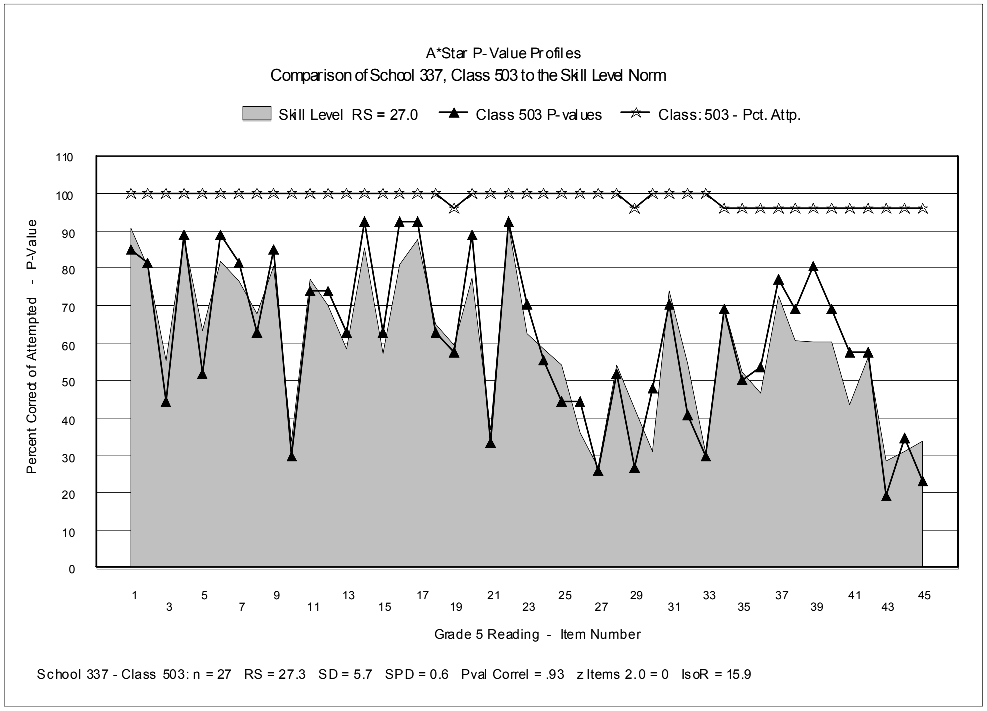
School 337 – Class 501: This class exhibits a dramatic deviation from the response pattern norm for its skill level, while Classes 502 and 503 closely follow their respective norms. When we look within Class 501, we find 11 students whose responses most substantially contribute to the irregularity of the full class response pattern over the items 26 – 33. The response pattern for the remaining 12 students is much closer to the norm over these items, but nevertheless contributes to the irregularity at other test items and test sections. The class responses clearly fail to demonstrate the independent student work necessary for meaningful measurement and do indicate the coordinated, improper involvement of their proctor.



School 337 – Class 502: The achievement level of this class is approximately one standard deviation above the norm and has good agreement with its skill level norm (r = .89).



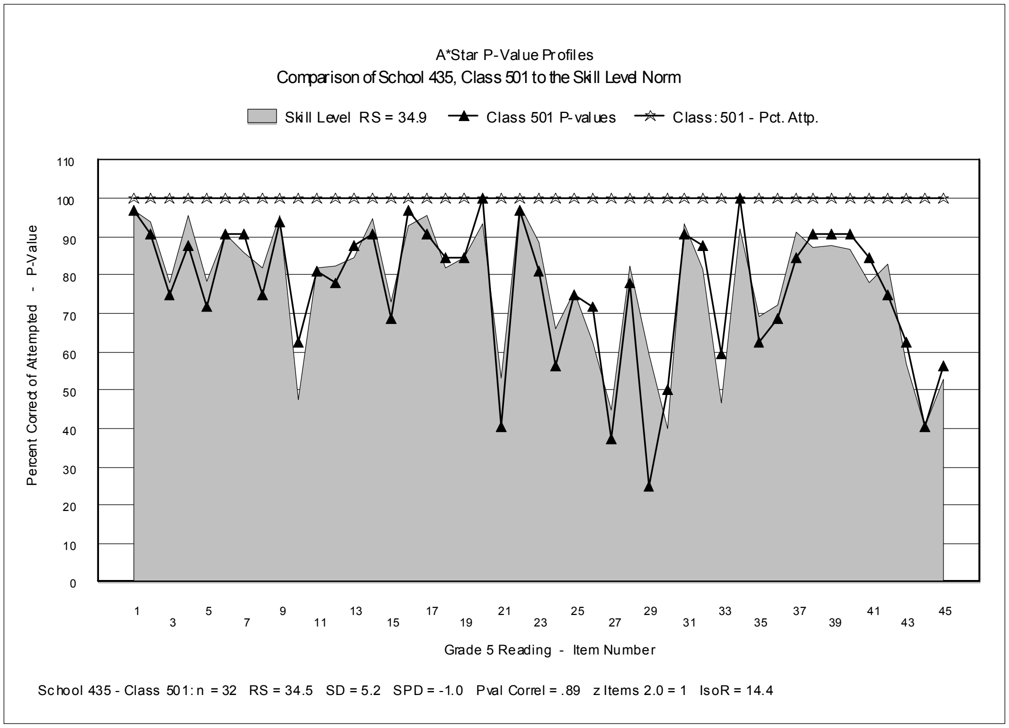
School 337 – Class 503: This class has good agreement with the norm (r = .93). The elevation at the end of the test is unexpected, but insufficient to draw criticism.



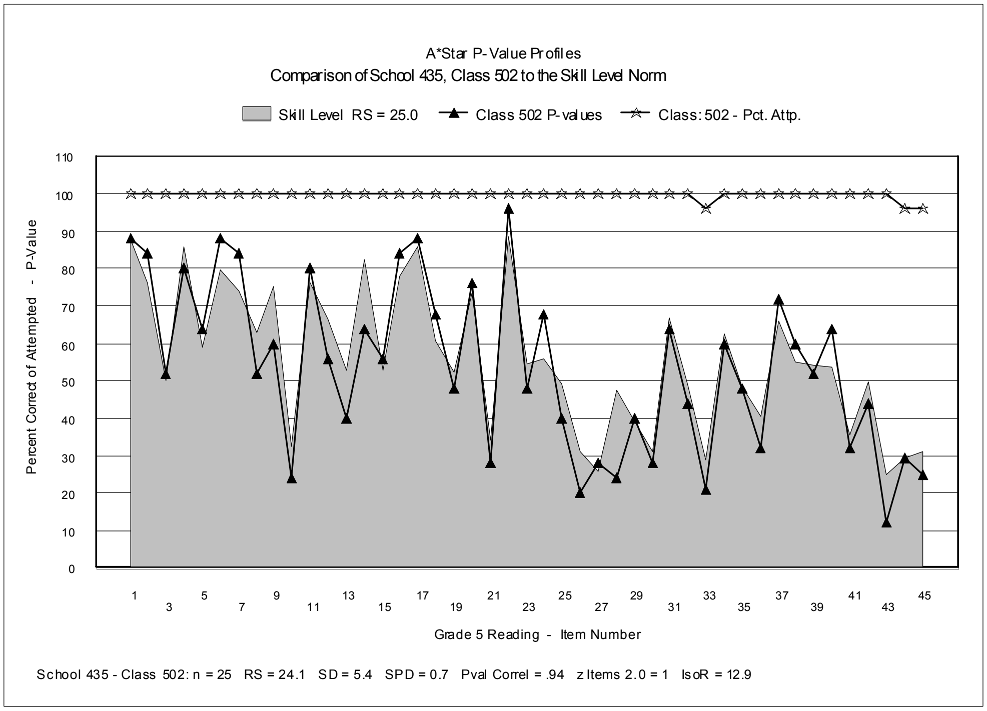
School 435

School 435 has 5 classes ranging in class average score from 22.9 to 34.5 and class size from 22 to 31. All 5 classes demonstrate an excellent agreement with their respective skill level norms. Essentially all students in each class responded to all test items. The A*Star experience suggests that this high level of test completion is not likely to occur without teacher encouragement to guess to complete answers that would otherwise be left blank. The teacher/proctors at this school were able to manage this task without significantly interrupting the measurement provided by the test.

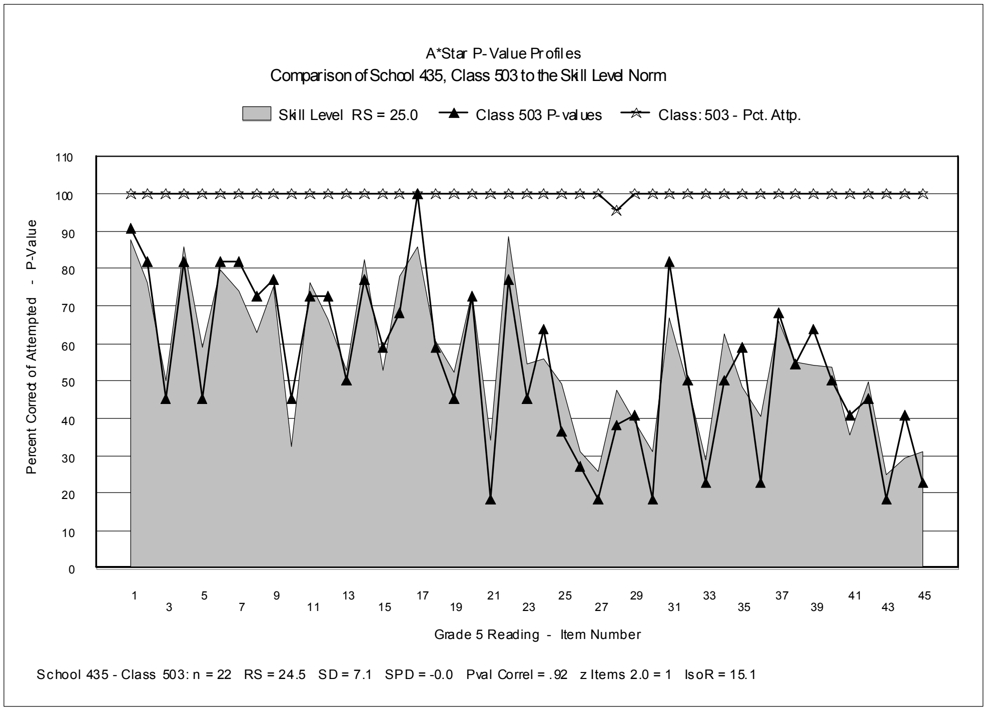
School 435 – Class 501: This class is above average in achievement and has good agreement with its skill level norm (r = .89).



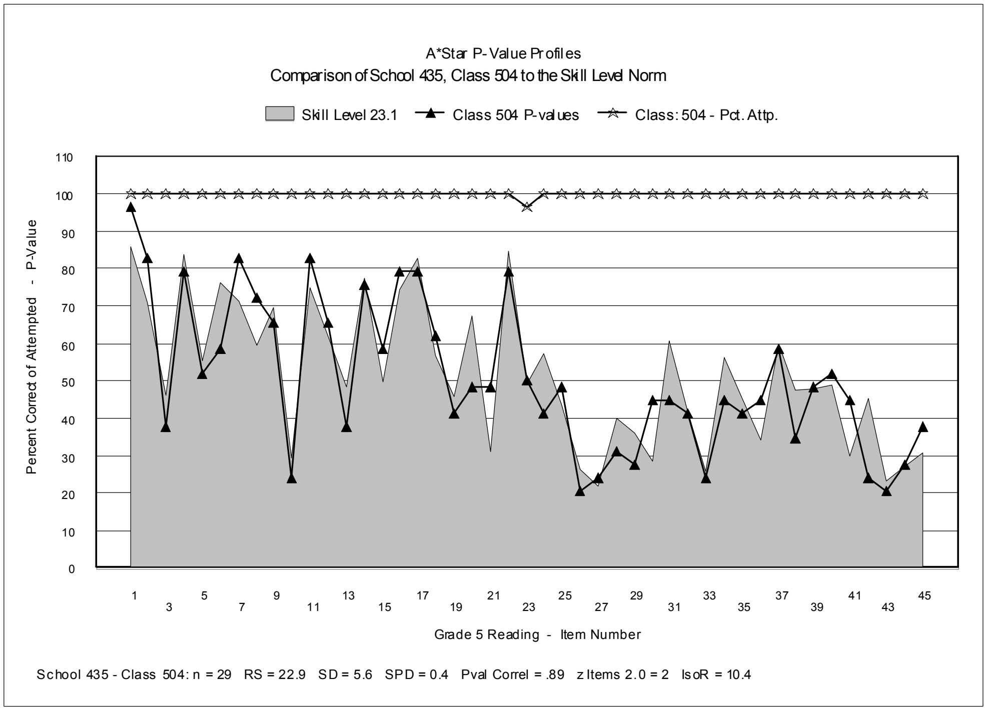
School 435 – Class 502: This class is well below the school district average, yet has excellent agreement with its skill level norm (r = .94).



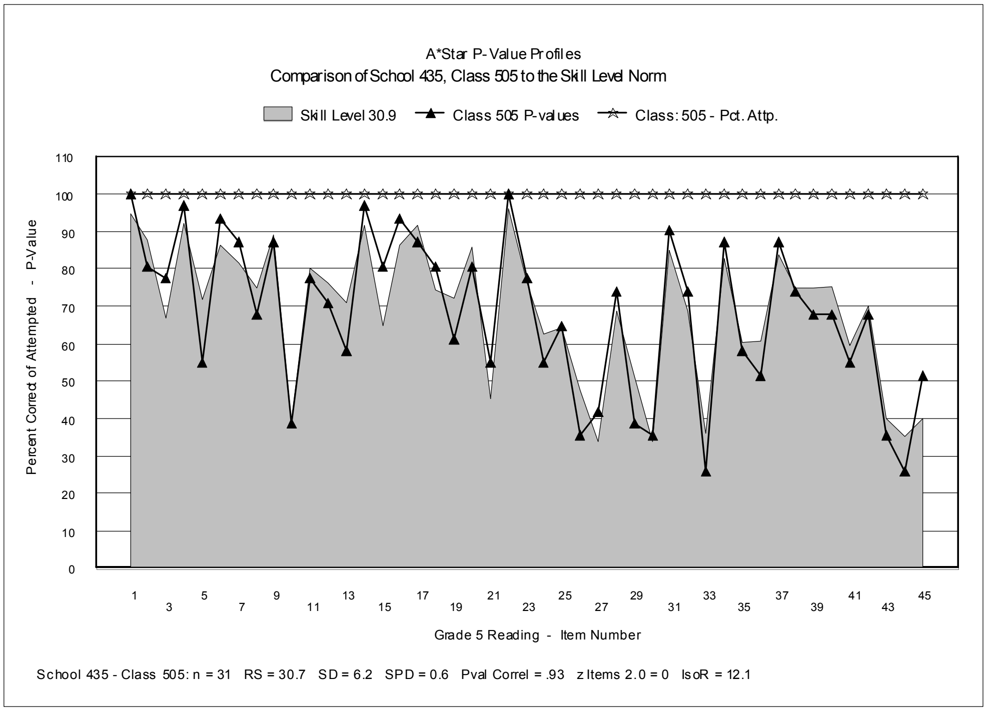
School 435 – Class 503: This class is almost a mirror image of Class 502, below the school district average in achievement, yet with excellent agreement with its skill level norm (r = .92).



School 435 – Class 504: The achievement level of this class is approximately one standard deviation below the school district average, yet has good agreement with the norm (r = .89)



School 435 – Class 505: The achievement of this class is slightly above the school district average and has excellent agreement with its skill level norm (r = .93)

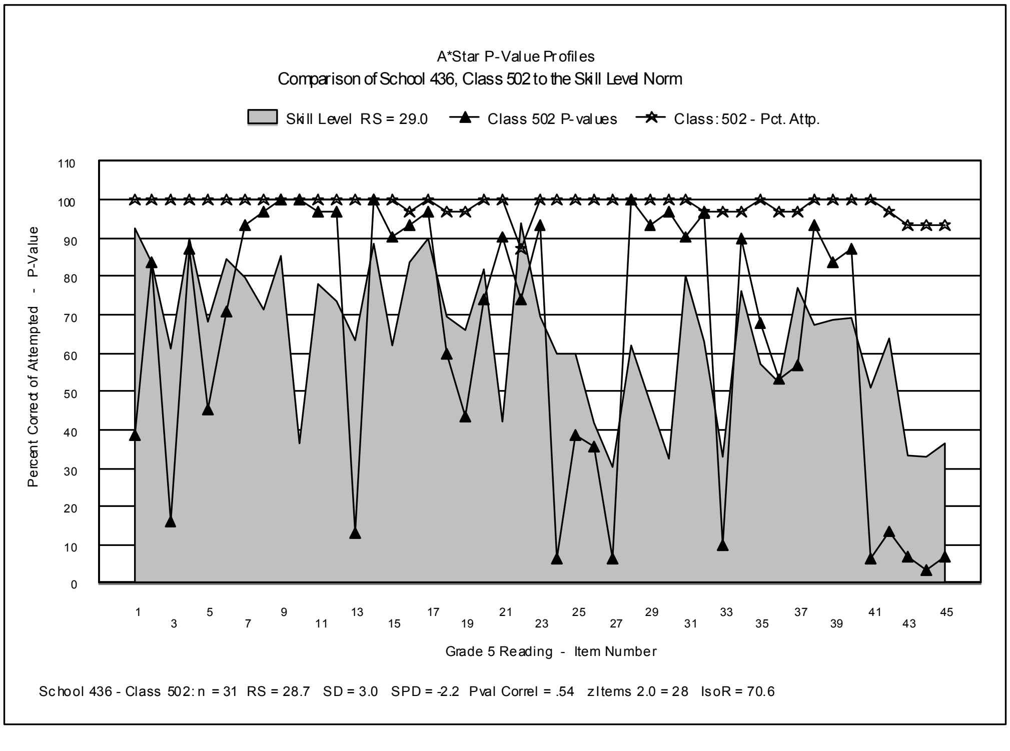


School 436

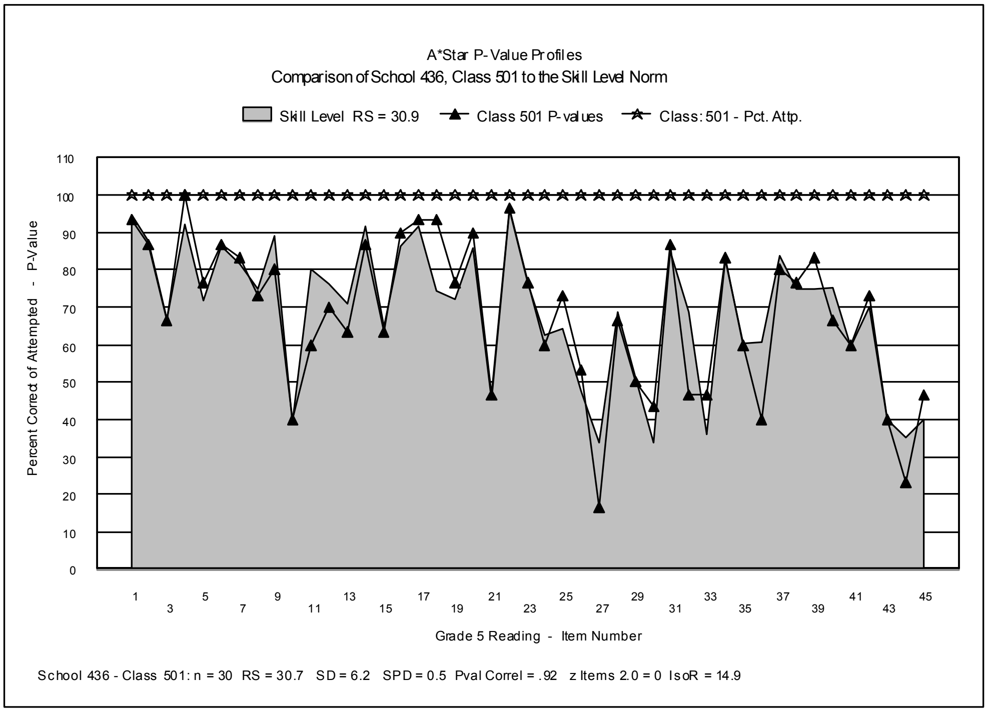
School 436 has 5 classes ranging in test score average from 15.6 to 30.7 – from the second lowest class average score in the school district to the approximately the school district average. Classes 502 and 506 have each received an A*Star Audit score of 3, indicating severe irregularity. The response pattern of Class 502 is representative of a concerted proctor intervention to control the students' test item responses. The response pattern of Class 506 is representative of a substantial confusion in the test administration, perhaps caused by the proctor's efforts to hurry low skilled students along or by the proctor offering suggestions on how to guess. In both classes, the test measurement has become invalidated by the proctors' failure to follow and maintain the proper test administration procedures.

School 436 – Class 502: Note, Class 502 is presented here out of order.

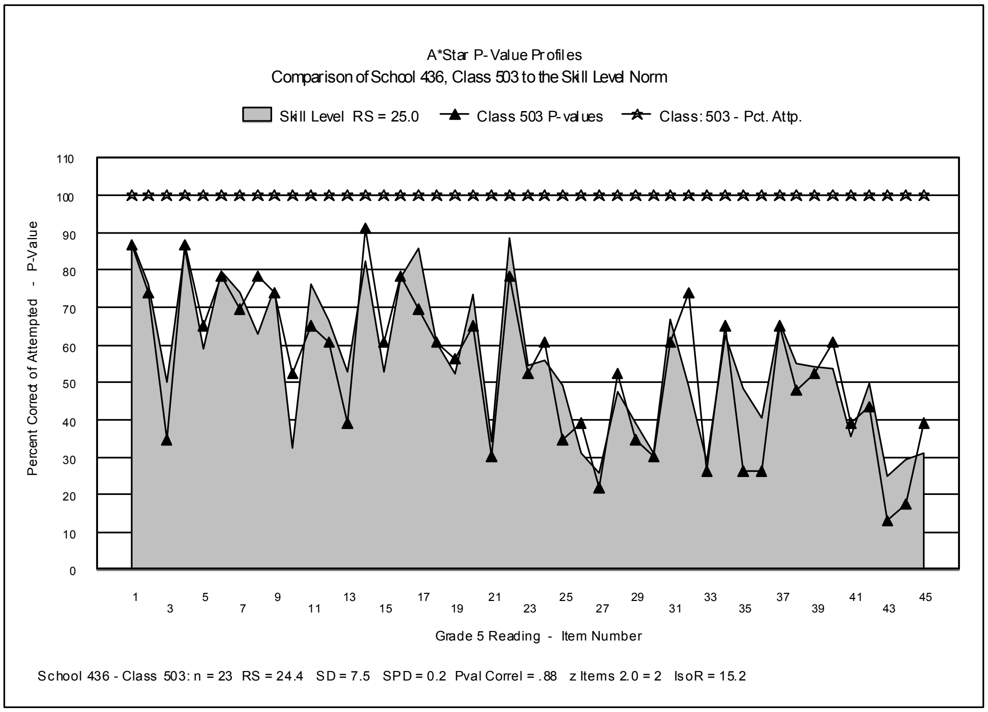
Class 502 has 28 of 31 students involved in the response pattern irregularity. It is quite likely that the three remaining students where not tested in the classroom, but were transferred out after the answer documents were pre-coded. There is a common answer pattern in this class and a review of the answer documents finds 145 erasures from some other answer to agree with this pattern. The character of the erasures indicates that they were made by the students (varying in neatness and darkness) under the direction of the proctor. There was one student in the class with a perfect match to the pattern, but no erasures. The high volume of erasures was not caught by the school district's routine erasure analysis.



School 436 – Class 501: This class is approximately at the school district average and has excellent agreement with its skill level norm (r = .92)

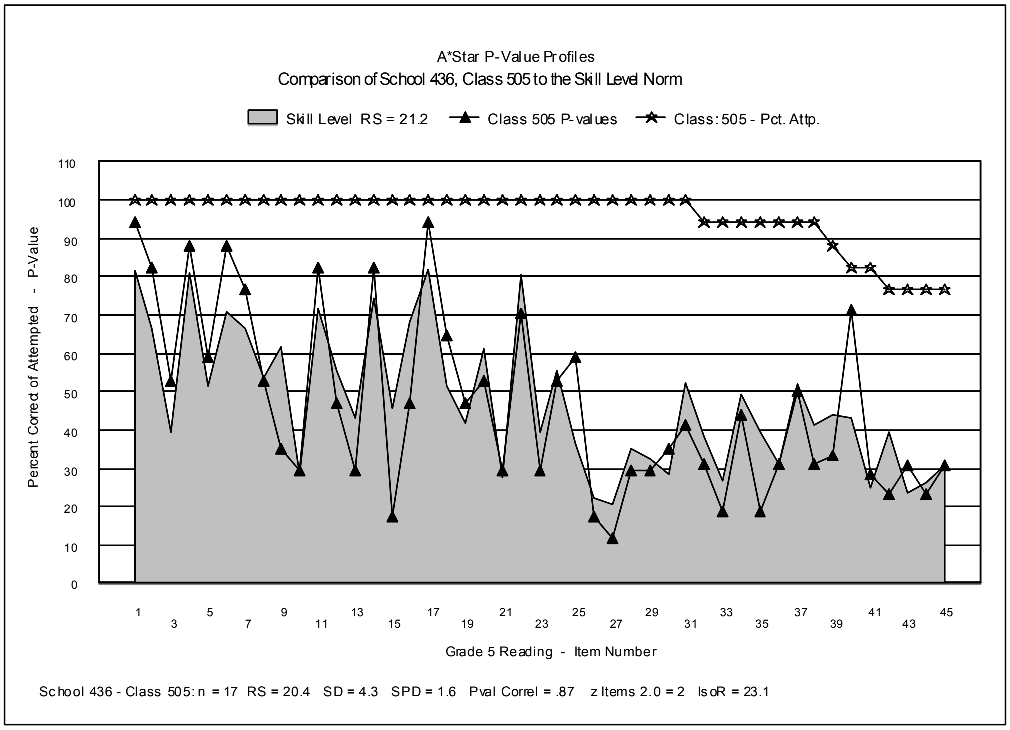


School 436 – Class 503: This class is well below the school district average (at the same score level as Classes 502 & 503 in School 435), yet has good agreement with its norm (r = .88)



School 436 – Class 505:

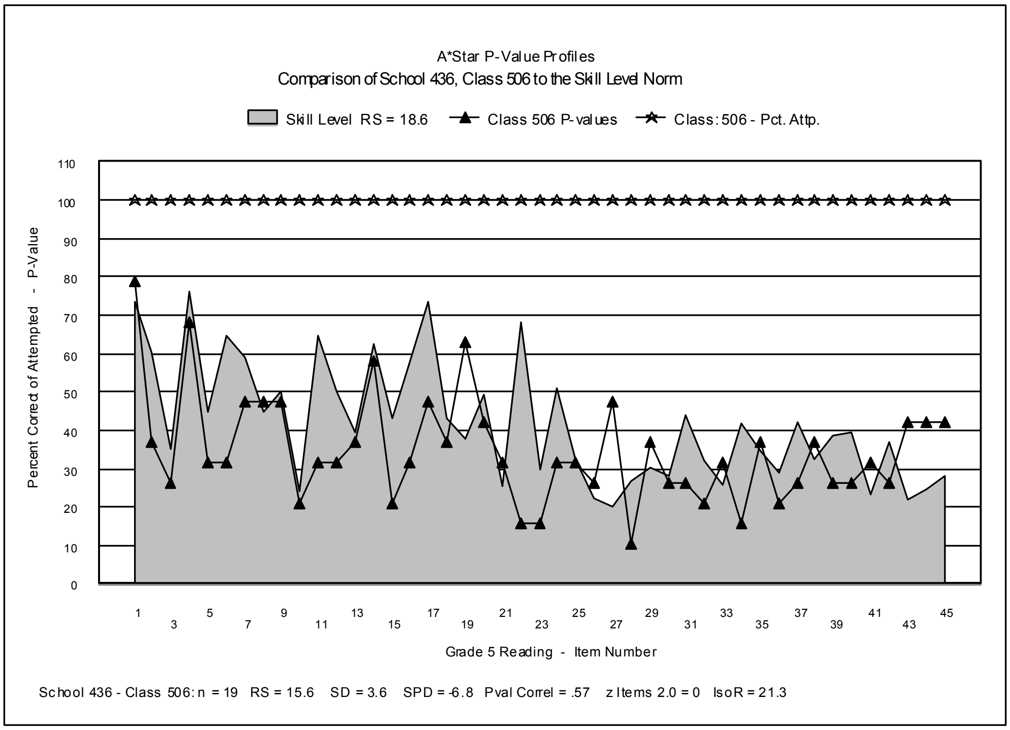
Class 505 has received an A*Star Audit score of 1, indicating a modestly irregular response pattern. Statistically, this is due to the elevated number of items with higher than expected p-values and the concentration of these items in one test section. A review of the class graph, however, suggests that this condition is due to the class beginning the test session working carefully, then urged to hurry and then pretty much left to their own. The careful work resulted in the higher early p-values and the hurry-up resulted in lower p-values that followed. It appears that the class was encouraged to guess to complete more answers, but not pressured to do so. The number of items left blank, indicated by the declining percent attempted, suggests a lack of pressure from the teacher/proctor.



School 436 – Class 506:

Class 506 has received an A*Star Audit score of 3 due to the extraordinarily low p-value correlation with the norm (r = .57) and the large differential between its early test and late test performance (6.8). The class graph indicates a promising early start over the first 4 or 5 items, then an interruption, then a few more representative items, and then essential randomness. The class average p-value over the last 25 items of the test was 29.05%, only slightly better than random guessing. There can be little doubt that many of these students would achieve higher scores under a proper test administration.

The teacher/proctor responsible for Class 506 is not likely to have made an intentional effort to assist the students with their answers, but rather to have inappropriately disrupted their independent work behavior to make certain that all answer blanks were filled in.



School 436 - Class 506: n = 19 RS = 15.6 SD = 3.6 SPD = -6.8 Pval Correl = .57 z Items 2.0 = 0 IsoR = 21.3

This page is intentionally left blank.