Case Study: Using Statistical Process Control (SPC) Charts to Analyze Standardized Test Scores[1]

Hallsville Independent School District (ISD), Hallsville, Texas

James F. Leonard, Consultant

A few years ago, I was provided data on Hallsville ISD's Texas Assessment of Academic Skills (TAAS) test results by Dr. Elizabeth Clark, Assistant Superintendent for Educational Operations. Elizabeth requested that I use statistical process control (SPC) charts to analyze the test scores and then submit recommendations for improvement. The reading and math charts are shown on Figures A and B, respectively, and the following comments are offered as a supplement to the data, control limits and patterns illustrated on the charts.

Introduction

The whole idea behind statistical process control charts – and the intermediate statistics in general – is the theory that variation is the enemy. In schools, this theory is based on the following very simple observation: Students already know or don't know the material by the time they sit down to take the TAAS test. The test itself adds no value. The enemy is the variation and sources of variation in learning and achievement that are swirling in and around the complex, dynamic classroom teaching and learning process – and students can learn no better than the process allows!

Of course, even though we view variation as the enemy, there will always be variation; between teachers, between students, between grade levels, between schools, and so on. The question to be addressed is, What is the variation trying to tell us about the teaching and learning process and about the people (teachers and students) who work therein?[2]

For example, Hallsville ISD's data exhibit variation in TAAS scores between and among grades 3 through 8 and 10 over a four-year period. They illustrate the tested outcomes from the K-10 components of the Hallsville ISD system. Control charts are used to determine what type of variation is present: common cause variation from within the teaching and learning process or special cause variation from outside the process. They will also indicate whether the grade levels' test results are merely different or *significantly* different.

If the data plot in a random pattern over the four-year period, this would indicate that the process is stable or in statistical control. In other words, the outcomes would be different but not significantly different. Such a state of stability would indicate that the K-10 components of the Hallsville ISD system, as measured by TAAS, are under the influence of common causes of variation only from within the teaching and learning processes. Outcomes from the stable process would be the results of Hallsville's curriculum design and content, texts and supplementary materials, teachers and staff, teaching and learning methods, technology, the test itself and other sources of common cause (or random) variation.

Intermediate, analytic statistical methods also help to make the effects of change clearly measurable. Ishikawa viewed control charts as means to see what changes in data occur over time, as well as the impact of various factors in the process that may change over time.[3] How can we tell if program changes have had the desired effect? The use of control charts helps to determine whether TAAS scores are moving not merely up, but up significantly. Significant (non-random) variation in the patterns of the data would indicate that the process has indeed changed (hopefully for the better!). If the pattern remains random, however, this would indicate that the process was not improved, or that changes have not had a significant effect on outcomes.

In conclusion, the purpose of applying the intermediate, analytic statistical process control techniques is to find out what type of variation we're up against: common cause or special cause variation. Another way to view the effort would be to determine the source of the enemy: common cause variation from *within* the process or special cause variation from *outside* the process. Knowing what type of variation is present helps school and district leaders to select the appropriate strategy for corrective action.

In the case of common cause variation, leaders could form a cross-functional team of people from different groups involved in or affected by the process: classroom teachers, specialists, administrator(s), parent(s) and, in the case of high school programs, student(s). The team would be directed to identify the major causes of low levels of achievement, then submit recommendations – backed up by data – for changes we can make to the process and programs to improve future achievement levels.

In most cases in manufacturing, evidence of special cause variation is bad news. It's an indication that some special cause from outside the process has thrown the process out of control or into a state of chaos. In such cases, leaders would direct people to take or recommend special action to find, remove and prevent the recurrence of the special cause. If such a special cause is not removed and prevented, it will come screaming in without warning in the future to frustrate any efforts to control or improve the process. No amount of work on the standard process or program will address a special cause, because by its very nature special cause variation comes from *outside* the process.

In some cases, however, evidence of special cause variation can be *good* news! In a school district, let's say the latest round of standardized test scores plot above the upper control limit (UCL); or a significant (non-random) shift above the central line (CL); or a significant (non-random) upward trend. Such patterns would indicate that efforts to improve student achievement have had a significant effect. Test scores have not gone up at random; they've gone up significantly as a result of the changes to the teaching and learning process; they came out of a different (better) process than the one that produced previous years' test results.

Control Charts Selected

To analyze Hallsville's TAAS scores, which report the percentage of scores above grade level or mastery, I selected the *p* chart to evaluate the reading and math program outcomes. In most manufacturing applications of the *p* chart, it is used to analyze the percent or proportion of defective outcomes. In the case of TAAS scores, we are analyzing the percent or proportion *effective*, because we're dealing with scores above mastery. The *p* chart is particularly useful in a school setting for two reasons.

First, unlike many manufacturing processes, educators can't "stack the deck" to process lots or batches of 50 units at a time every time. Different classes have different numbers of students. The *p* chart, however, allows one to take the data as they come, with no need to manipulate the measurements into standard samples or units of equal size.

Second, the *p* chart automatically factors in how the same number of incidents will have a very different impact as a percentage on groups of different sizes. For example, in a class of ten students, if one is absent we have a percent absenteeism rate of ten percent. In a class of twenty students, one absence produces a percent absenteeism rate of only five percent.

The same number of incidents has a very different percentage impact on small classes versus large classes. This is one reason small schools or classes live in fear when district report cards of TAAS scores are published by the central office. If you're a small class, you don't need many kids to score below mastery to look *terrible* as a percentage!

By the same token, over the four-year period reflected in the Hallsville study, some grades had as few as 190 students while others had as many as 280 students. It would be neither fair nor rational to compare such different-sized groups on a strict percentage basis (even though administrators and state auditors do it all the time). The *p* chart automatically factors in the reality of higher impacts as percentages on smaller groups.

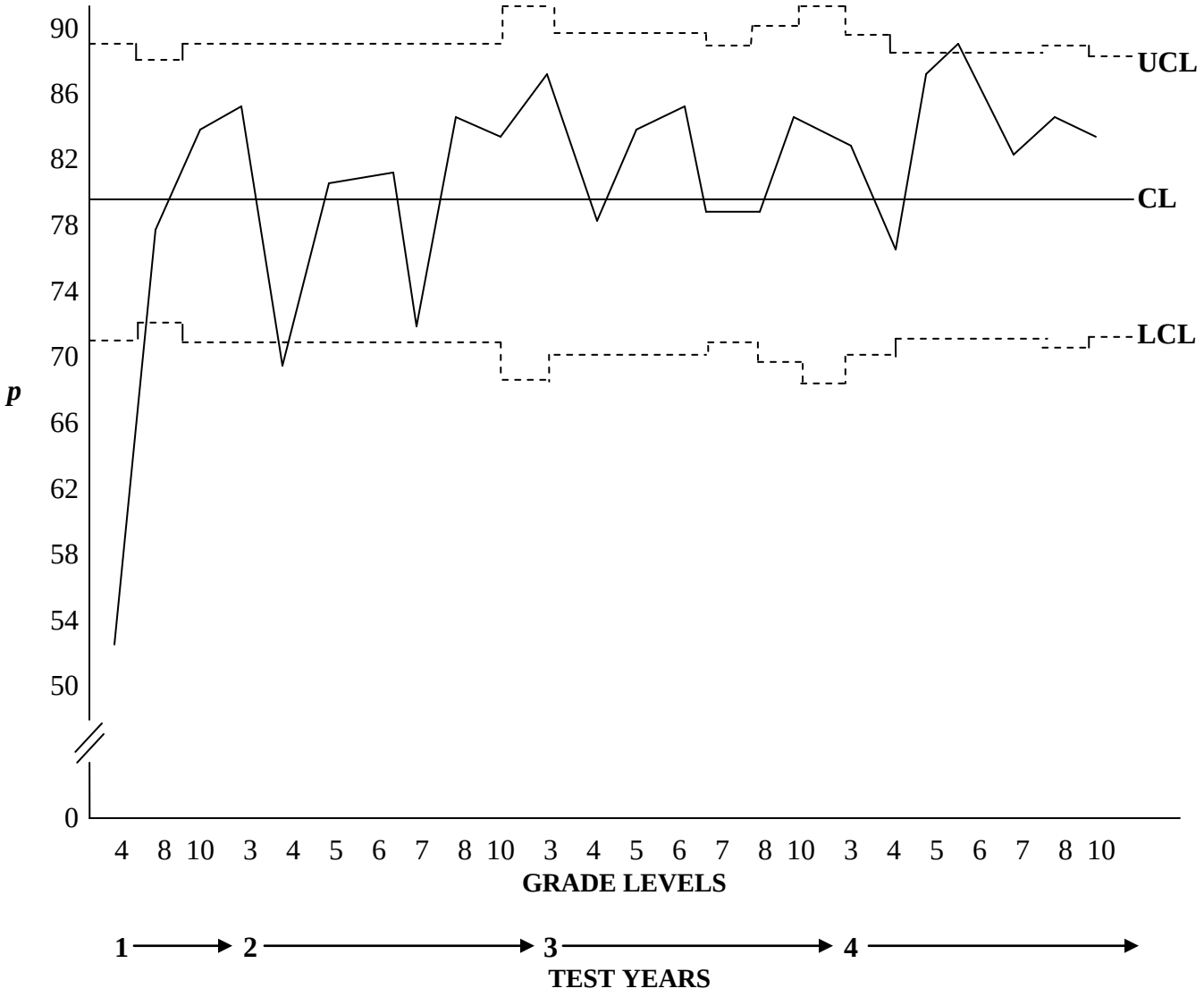
Like all control charts, the *p* chart has upper and lower control limits (UCL, LCL) of controlled, random, common cause variation. In this case, those limits were derived from all of the district's TAAS scores for grades 3 through 8 and 10 over the four-year period (24 measurements representing hundreds of student test scores). If Hallsville's ISD scores plotted in a random pattern over the four-year period, with no points falling outside the UCL and LCL, this would indicate that the K-10 teaching and learning process was stable. It would further indicate that any program changes or improvement efforts adopted over the same four-year period did not have a significant effect on outcomes.

On the other hand, if a non-random pattern emerged – especially a significant upward trend – this would indicate that the district's efforts to improve their programs did have a significant effect. The improvements would have produced outcomes that were not only higher than past TAAS scores, but *significantly* higher.

Figure A

Figure A below illustrates the *p* chart for the reading TAAS scores for grades 3 through 8 and 10 for the four-year period. Note that in the first two years, two scores plotted below the lower control limit (LCL). In comparison to the limits for the K-10 components of the Hallsville system (defined by the UCL and LCL), these scores are not only low, they're significantly low. However, in the last two test years, no scores fell significantly low and one score plotted significantly high (above the LCL) in year four.

Figure A. *p* Chart: Hallsville ISD Reading TAAS Scores
Percent *READING* Scores Above Mastery *(p)*



Analyzing the data in this form indicates that Hallsville ISD has accomplished significant improvements in its K-10 reading teaching and learning processes over the four year period. High scores are significantly high, and there are no longer any significantly low scores. Efforts to improve reading programs have succeeded with clear and measurable effects.

Figure B. *p* Chart: Hallsville ISD Reading TAAS Scores
Percent *MATH* Scores Above Mastery *(p)*

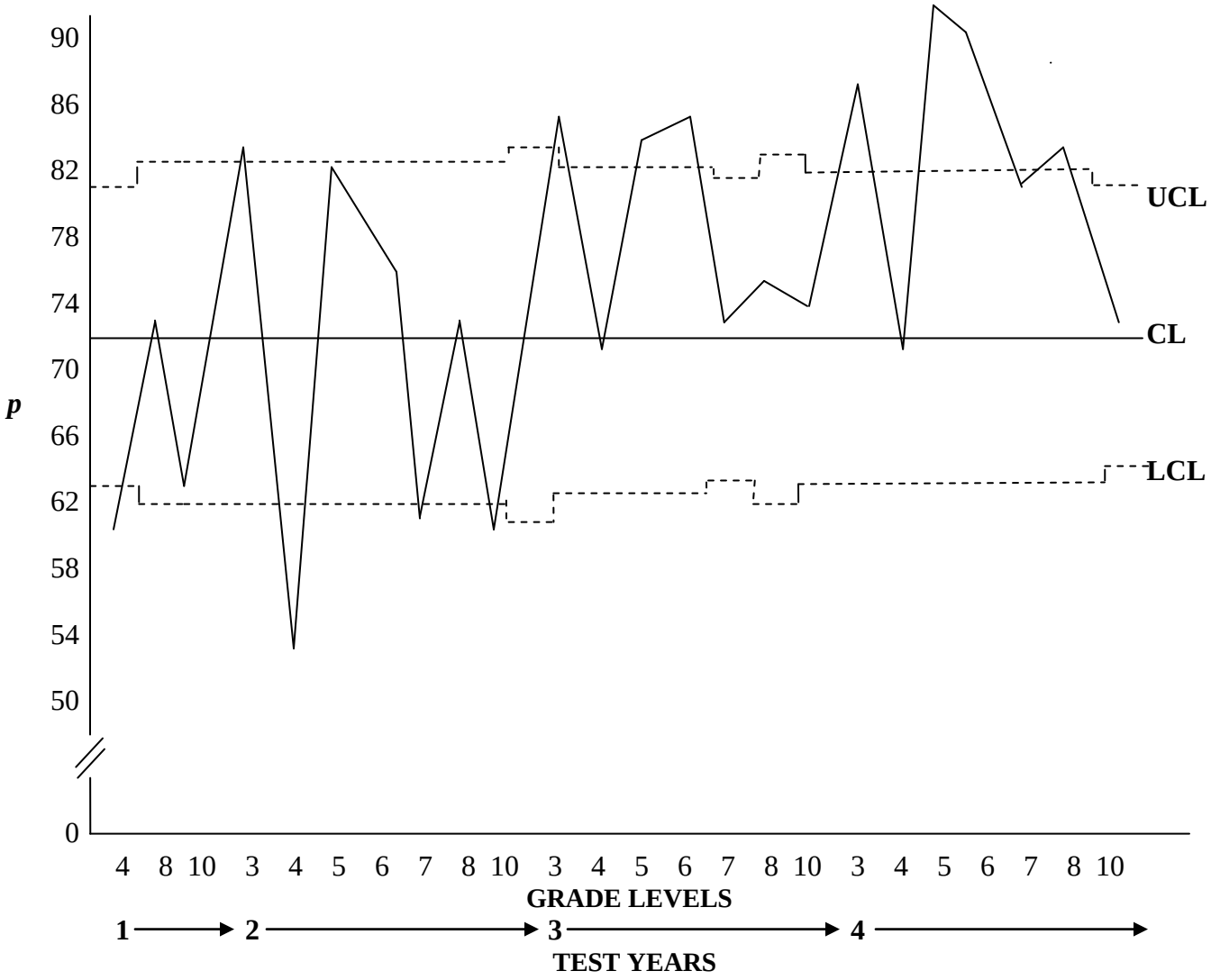


Figure B

As in the case of the reading scores, the *p* chart for the district's math scores illustrates significant improvements over the four test years. In the first two years, four scores fell not only low, but significantly low. In the last two years, we find no scores falling below the LCL and *seven* scores plot significantly high (above the UCL). This indicates that changes made in the past few years have had a statistically significant effect on the K-10 outcomes, as measured by the math TAAS test.

A Word of Caution

In the face of the wonderful improvements revealed on Figures A and B, it must be stressed that district leaders can't have it both ways. They cannot celebrate the significant improvements accomplished in Hallsville ISD, then turn around and hammer fourth-grade teachers for their low TAAS scores! The fourth grade is one component of a system that's showing significant improvements. However, district leaders must be prepared to respond to questions or complaints about the low (though no longer significantly low) fourth grade scores. Two potential observations merit discussion.

1. In year 2, 85 percent of the district's third graders scored above mastery on the reading TAAS test (see Figure A). The following year, only 78 percent of the same students scored above mastery in the fourth grade. In year 4, 86 percent of the same students scored above mastery in the fifth grade. One might wonder, "What's going on in the fourth grade? The same kids score a lot lower there than when they're tested in the third and fifth grades."

   I have no response to such a question, except to note that all three outcomes (year 2 third grade score, year 3 fourth grade score, and year four fifth grade score) fall within the UCL and LCL on the *p* chart in Attachment A. Therefore, those outcomes may be different, but they're not significantly different. Granted, the fourth-grade score may be unacceptably low, but it was produced by the same system that produced the third and fifth grade scores in test years 2 and 4, respectively. If district leaders aren't happy with these outcomes, they must work on the K-10 system, striving for K-10 components of the Hallsville ISD system that work well together – as opposed to perfect, individual components in grades 3 through 8 and 10.

2. In year 2, 83 percent of the third graders scored above mastery in math (see Figure B). That percentage dropped to 71 percent in the fourth grade in year 3, then rose to 90 percent for the same students in the fifth grade when they were tested in year 4. Once again, one might note these differences and be tempted to criticize or to attack the fourth grade teachers.

As illustrated on the *p* chart in Figure B, the year 2 third grade score and the year 4 fifth grade score plot above the UCL, while the year 3 fourth grade score plots within the control limits. The reason for these significant differences, however, may not be bad teachers or a poor math program in fourth grade.

An alternative explanation would be that the fourth grade teachers are doing a *great* job preparing their students for the fifth grade! The third grade teachers may be "drilling and killing" on the third-grade TAAS test, thereby failing to adequately prepare students for the fourth grade math program and learning experience.

However, one cannot draw any valid conclusions about the year-to-year variation in TAAS scores for the fourth grade (or any other discrete component of the Hallsville ISD system) unless we have at our disposal *twenty years* of fourth-grade TAAS scores! In order to generate valid control limits, one must plot at least 20-25 measurements on a control chart. (You can, however, test a trend with just eight years of fourth-grade scores.)

Therefore, it is recommended that district leaders remain focused on improving Hallsville ISD's K-12 *system*, and not trying to focus on any one individual grade level component of that system. That would result in grade levels striving to become perfect, individual components. If that were to happen, the significant systemic improvements evident on Attachments A and B would not and could not continue. The children of Hallsville, Texas, need K-12 teaching and learning components that *work well together* – not perfect, individual components.

Recommendations

To continue the systemic improvements discovered in this analysis, it is recommended that district leaders continue, build upon and expand the efforts and programs that generated the improvements in the first place. These initiatives include:

- Building upon the Community of Learners model
- Cross-building, cross-grade-level, cross-functional project teams
- Application of statistical methods, with great emphasis on the basic tools
- Remaining data-driven, especially in efforts to correlate program changes to key measures that indicate the effect on students and student achievement. Such key measures will and must include assessments other that the annual TAAS tests, because a once-a-year answer to the question, "How are we doing?" is not robust enough to drive continuous improvement.

As it relates to cross-grade-level teams, the district should continue to include parents, local business partners and other members of the Community of Learners in selected improvement

projects. However, additional emphasis must be placed on expanding cross-grade-level cooperation and collaboration among *teachers*. Schmoker noted that "teachers, on the front line in the battle for school improvement, are working in isolated environments that cut the lifeline of useful information. Such isolation thwarts them in developing common solutions through dialogue."[4] If significant and continuing improvements in student achievement are important to district leaders, they must eliminate the isolation and remaining barriers between teachers, grade levels and buildings.

In closing, it was an honor and a pleasure to work with Jim Dunlap, Superintendent of Hallsville ISD, Elizabeth Clark and other Hallsville educators. I hope that in some small way I made a contribution to the improvements reported in this case, and I am very grateful for Elizabeth's kind feedback, as follows:

> "Under Jim Leonard's guidance, we have applied his tools, principles and concepts to achieve a number of impressive improvements. Local business partners are reporting a much higher level of respect for our district; new behaviors are evident among administrators and teaching staff; and student test scores have shown statistically significant improvement in just a four-year period." — *Dr. Elizabeth Clark, Deputy Superintendent, Hallsville Independent School District*

Notes

[1]J.F. Leonard, the *New Philosophy for K-12 Education: A Deming Framework for Improving America's Schools,* ASQ Quality Press, Milwaukee, WI (1996), pp. 129, 134, 136-137 and 162.

[2]W.E. Deming, *The New Economics for Industry, Government, Education,* MIT Center for Advanced Educational Services, Cambridge, MA (1993), p. 101.

[3]K. Ishikawa, *Guide to Quality Control,* Asian Productivity Organization, Tokyo, Japan (1986), pp. 61-62.

[4]M. Schmoker, *Results: The Key to Continuous School Improvement,* ASCD, Alexandria, VA (1996), p. 10.