

The Numbers Game: It Is Not Just the Final Score; It REALLY IS About How You Play The Game!

Deborah Dusseau, Ed D, David Hurst, PhD, and David Bitter, Ed D

This article appeared in the September 2003 issue of *Principal Leadership* and is posted with permission from *Principal Leadership*. *Principal Leadership* is a publication of the National Association of Secondary Schools Principals (NASSP). For more information concerning NASSP services and/or programs, please call (703) 860-0200, or visit www.principals.org.

Introduction

Assessment is much more than a game, and the game really is about more than the final score. Educators who are committed to making a difference (as opposed to documenting a difference), will make great strides by considering critical elements of the game before they engage in school improvement. This article addresses research design, assessment program design, and interpretation of results.

What Game Are We Going to Play?

An early consideration in the assessment game is to be clear about what you plan to measure. Before a valid and productive assessment program is developed, the purpose(s) of assessment must be determined. You need to know what game you are playing! You might go to the game well equipped with your "Louisville Slugger", but it won't do you any good if the game is volleyball!

The Cohort Growth Game

The requirements of No Child Left Behind [NCLB] and many state mandated assessment plans are focused on student growth. This has resulted in an intense focus on longitudinal measures and analysis. Longitudinal analysis, a comparison of assessment results from a student group as it progresses from year to year, is often referred to as cohort growth. One might say, "Let's track the growth of students in

the Class of 2012.” The results would provide a comparison of the performance of these students as fourth graders in 2003-04 with their performance as fifth graders in 2004-05, as sixth graders in 2005-06, and so forth until they graduate in 2012. This would allow a school to analyze (among other things) the amount of growth from year to year (Adequate Yearly Progress), to identify strengths and weakness in certain skill areas within the group, and to evaluate whether the students, as a group, are meeting the standards and benchmarks established in the curriculum. The reliability and usefulness of cohort growth measures can be compromised by a number of factors including high mobility of students, changes in the curriculum, changes in the assessments, and changes in the required achievement levels or evaluative norms. Since student populations have become very mobile, the most influential factor is probably the change in the cohort population. The composition of the fourth grade cohort in 2003-04 will probably change dramatically by the time the group reaches twelfth grade (Fowler-Finn, 2001). These concerns can be overcome if a school has sufficient data analysis and statistical sophistication.

The Program Growth Game

Assessing program growth yields different information and may be easier for educators to design, implement, and understand at the school level. Educators often complain that many of the outcomes for which they are held accountable are beyond their control. Educators don't control the ability levels, numbers, family backgrounds, level of parental support, or native languages of the students who come to their classrooms. In other words, educators do not control the cohort. Conversely, what happens **within** the classroom—the educational program-- is largely under the

control and influence of educators. The educational experiences the students engage in, the teaching techniques used, the timing and pace of the curriculum, and how learning is practiced and reinforced are largely controlled by the instructor. Program growth focuses on the impact of these variables.

Measuring program growth allows educators to compare whether they are doing a better job for students now (after initiating changes, new programs, interventions, and/or improvement activities) than they were doing in the past. Program growth is measured by comparing assessment results at the same grade level for multiple years. For example, if a new reading program were initiated in an elementary school, a school could evaluate the effectiveness of the new program by comparing overall reading scores for the fourth grade students in 2003 (before the new program) with fourth graders' scores in 2004, 2005, and 2006. There are confounding issues with the results of program growth measures. Some common concerns are the consistency of implementation of new practices and related staff development issues. However, these concerns typically can be addressed by educators with moderate familiarity about program growth research design.

School accredited by the North Central Association Commission on Accreditation and School Improvement [NCA CASI] have been focusing on program growth measures for the last fifteen years as part of performance based accreditation. This focus is based on a belief that effective schools engage in a continuous school improvement process that gathers information about the needs of the students. Further, schools must plan and implement interventions to address the needs, evaluate the effectiveness of the interventions, institutionalize those

interventions that improve student performance, and discard those that were ineffective or reduced performance. While implementing this protocol, NCA CASI has identified common impediments to assessment design which, if not addressed, will compromise the validity of student performance results.

What are the Rules of the Game?

Dr. Robert L. Armstrong (Armstrong 2000), a professor emeritus in mathematics at Arizona State University developed “rules of the game” for assessment design by categorizing a number of assessment practices into Green Light and Red Light conditions. Armstrong called the desirable assessment practices “Green Light Conditions” and those to be avoided “Red Light Conditions.” Greene, Winters, and Forster (2003) substantiate some of these conditions, as well.

Green Light Conditions

Five basic conditions must all be met before an assessment is a valid measure for program growth. In other words, the condition must be “green” before one can proceed. These conditions include:

1. The pretest (baseline) and posttest instruments must be the same or psychometrically equivalent for any given assessment. This is a basic tenet of pre-and-post testing. Unless the instruments are the same or equivalent, the scores will have no meaning.
2. Pretest and posttest assessments are conducted on the same grade level(s) of students. Testing is done at the same grade level because program assessment compares the achievement of students taught in the former way to the performance of the students taught the same content in a new or improved way.

3. Pretest and posttest assessments are administered at the same time in the respective academic years. This helps ensure that the comparison of students happens at an equivalent point in both intellectual maturation and in the amount of time provided to learn the subject matter.
4. The pretest (baseline) assessment is administered close to the time that the implementation of the new programs or interventions will begin. If the baseline data are old, the effects of the changes or new programs become blurred.
5. The implementation period is sufficiently long for the new practices to have an authentic effect. It takes time for teachers to assimilate and optimally practice new skills and habits. If the baseline-posttest time period is not long enough, the new methods or program will not have been completely implemented. In addition, if the change is complex, a J-curve phenomenon (a decline in student performance) may result for a short time after implementation. Ideally, implementation should span two or three years and be measured at least annually.

Red Light Conditions

Six conditions might occur that could compromise the authenticity or usefulness of the assessment results. If one or more of these occurs, it creates a red light. A red light requires the people performing analysis to “stop” and assess the influence of the occurrence on the results of the assessments. Some of the Red Light Conditions affect the ability to convert the scores to standard units, which is essential to making statistical comparison between the baseline and posttest data. If a Red Light

Condition occurs, it doesn't mean the data are useless; it only means that they may not be valid for statistical analysis. The following conditions should be avoided.

1. Beware of using tallies as measures. A tally may simply count the number of occurrences such as the number of behavior incidents on the playground.

However, there is no fixed upper limit on such a number. On the other hand, the percent of students that were involved in such incidents does have an upper limit (100%) and would be statistically a more useful measure.

2. Beware of using posttest data that has no true baseline. For example, schools may want to measure the results of a new curriculum or program that has not been previously taught. The lack of baseline data prevents any assessment of program growth. However, the posttest data could be useful in cohort research design.

3. Beware of assessment results from non-random subsets of the population. Such groups might include ACT test takers, Advanced Placement test takers, or remedial or at-risk program students. While disaggregated data are useful in analyzing equity issues, they don't provide information about program growth for all students.

4. Beware of assessment results that come from a low or high ceiling assessment. Assessments that allow nearly all students to perform well, such as minimum competency tests, or assessments that sort out high achieving students, such as placement tests for honors programs, yield scores that do not provide discriminating information on the entire population.

5. Beware of assessment results that are expressed as grade equivalents. Grade equivalents are difficult to manipulate statistically since they don't represent

equal interval data. If other scores are available (and typically they are), they should be used.

6. Beware of assessment results that are expressed as stanines. Stanine scores are approximate values and represent a band of scores rather than a point. If one begins an analysis with data that are an approximate conversion, any assumptions made in further conversions or comparisons are compounded by the initial problems from the “soft” data.

Once the rules of the assessment game have been determined, then you can apply the right play book. The final step is to know what the score means.

How Do You Know If You Won?

After you have determined exactly what you want to measure and considered the rules for evaluating your assessments, how will you know if you are truly improving? Two major issues come to mind: 1) multiple assessments provide a more comprehensive picture of student performance and 2) interpretation of results from multiple assessments requires the use of specific statistical tools.

Student performance is under public scrutiny as never before. In the authors' experiences in working with schools, the major obstacle to improved student performance is not the will to improve, it is more often that practitioners are not trained to use data at the sophisticated level that is required by today's accountability standards. Bernhardt, in her book *Data Analysis for Comprehensive School Improvement*, advocates that “. . . multiple measures must be considered and used to understand the multifaceted world of school from the perspective of everyone involved. . .” (1998, p.13) Sergiovanni (2000) recommends multiple assessments that

reflect both state and local standards. Further, “. . . teachers and administrators need a thorough understanding of fundamental concepts and principles of *both* [emphasis added] classroom assessment and standardized testing to effectively apply these tools to improve student achievement” (McMillan, 2001, p. 2). The conclusion is that we must rely on more than one source of data to substantiate that student performance, in the overall sense, is improving.

If a person accepts the premise that multiple measures are preferable over a single assessment, then the “game” gets complicated. In comparing results, you could be dealing with assessments that use different metrics, different assessments that use the same metric, or assessments whose metrics cannot mathematically be manipulated. In all of these instances the results must be converted to a standard unit for comparison purposes.

If one assessment uses percentile scores as a metric and another uses raw scores, how can both assessment results be used? How do the results relate to one another? One relatively simple answer to this problem is to compare standard scores, or z scores. Simply defined; “a z score is expressed as units of standard deviation above or below the mean” (McMillan, p. 116). By determining the difference between pre-test z score and a post-test z score, we can measure the growth in standard deviations—a “common denominator” for statistics. This difference is known as effect size, or the magnitude of change (not to be confused with statistical significance, which comes from an entirely different can of statistical worms).

Because of the growing number and varied types of assessments and the increased focus on assessment data for accountability, researchers are turning to

standard scores and effect size as indicators of improvement. These tools allow comparisons to be made between and among disparate assessment scores. For example, in a recent study of high-stakes testing, Greene, Winters, and Forster (2003) utilized standard scores in part "because scores were reported in different ways (percentiles, scale scores, percent passing, etc.)..." (p. 4).

Calculating z scores and effect size is a relatively simple task. There are several software packages available that can calculate effect size simply and quickly. Most statistics textbooks provide tables that allow standard scores to be calculated with little stress. When using a table from a textbook, remember that the mathematical assumption is that the population you have selected is representative of a standard distribution. Rarely do data sets produce perfect bell-shaped curves. However, if you accept this assumption, standard scores can be calculated in a reliable manner, especially those resulting from standardized tests which represent large populations. Armstrong (2000) referred to this concept of using a given population as a subset of a larger population for the purposes of calculating effect size as the "adapted standard unit" [ASU]. Presenting data in terms of standard scores and effect size can reduce confusion and simplify the reporting process when multiple assessments are used.

Even after data are converted to standard units, a major question remains: how much growth is "good"? Some educators purport that equal increments of growth over 12 years is the definition of "good" growth and, at the end of that time, all students will be successful. We deem this view to be in conflict with what we know about human development; nonetheless, some quantification is needed to be

able to define “good” growth. Armstrong (2002) conducted a multi-year study of 600 NCA CASI accredited schools that used adapted standard scores to measure the results of their school improvement activities. Based on that study, Armstrong concluded that an effect size or standard unit growth of 0.3 was indicative of substantial growth, growth of .2 to .29 standard units was considered good, and growth of .1 to .19 standard units was worth mentioning. Declines in performance were defined by the same scale. Readers should realize that effect size (or standard unit growth) is a subjective term and should be defined in the context of the study.

Using the assumptions above, let us examine some sample data. Table 1 below is designed to present improvement (or the lack of improvement) in instruction over time. This table shows the results in adapted standard scores for three assessments that use different metrics.

Table 1: Demonstration of the Adapted Standard Unit (ASU) for Program Growth

Assessments that Measure the Same Improvement Goal	Type of Score	Pre-Test Score	Pre-Test ASU*	Post-Test Score	Post-Test ASU*	△
Nationally norm-referenced test	NCE	45	-0.13	49	-0.03	.10
State Assessment	Scale Score	257	.03	386	.74	.71
Local Criterion Referenced Test	Percent Correct	71	.55	84	.99	.44
Net Gain (Loss)						.42

Notes: The adapted standard unit (ASU) scores were calculated using values from a standard distribution table. The Net Gain (Loss) was calculated as an average of gain (loss) from all assessments listed for the improvement goal. The result, using Armstrong’s scale, would indicate the school had made substantial improvement toward the goal.

The graphic display below (Figure 1) is an extension of the use of adapted standard score method. The graph displays data that could be used to demonstrate cohort growth.

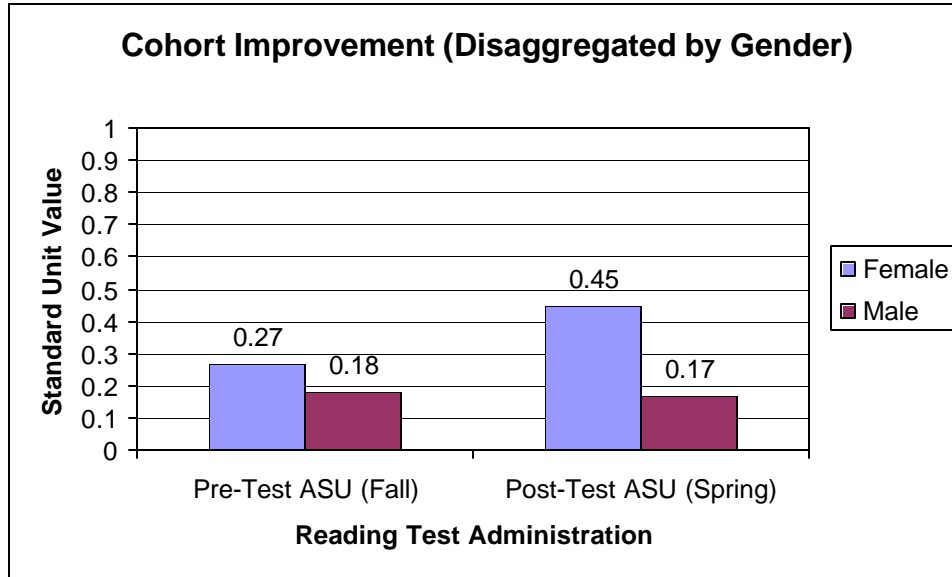


Figure 1: Values for each assessment are reported in standard units (standard deviations). Data indicate a .18 ASU* improvement in results for females, and a .01 ASU decline for males. The average improvement for all children is .09 ASU.
 *Adapted Standard Unit

Most statisticians would agree that mathematics is based on a set of assumptions, and results are valid only in light the assumptions used to calculate them. We believe that this method of standard score conversion and interpretation, based on the stated assumptions and NCA’s research results from 600 schools, provides schools with an accurate indicator of improvement.

Summary

The decision to document program growth, cohort growth, or both should be based on an honest representation of the data in context rather than being based on pressure to achieve a “winning score.” If the assessment design was compromised, the final score lacks meaning and credibility. Finally, the use of standard scores allows for accurate interpretation of multiple assessment results, thus reducing

dependency on single “high stakes” assessments. As in all sports, there is more to this assessment game than the final score.

References

- Armstrong, R. L. (2000). *Score Conversion Handbook: An NCA guide to conversion decisions*. North Central Association Commission on Accreditation and School Improvement: Tempe, AZ.
- Armstrong, R. L. (2002). Change in achievement in schools completing the NCA CASI school improvement process: A third update. *Journal of School Improvement*, Vol. 3 No 1, Spring 2002, North Central Association Commission on Accreditation and School Improvement: Tempe, AZ.
http://www.ncacasi.org/jsi/2002v3i1/change_achievement
- Bernhardt, V. L. (1998). *Data analysis for comprehensive school improvement*. Eye on Education: Larchmont, NY.
- Fowler-Finn, Thomas. (2001). *School Administrator*. Student stability vs. mobility. 58-7, August, p. 36-40. http://www.aasa.org/publications/sa/2001_08/fowler-finn.htm
- Greene, J. P, Winters, M. A., & Forster, G. (2003). Testing high stakes tests: Can we believe the results of accountability tests? Civic Report 33. Center for Civic Information at the Manhattan Institute: New York, NY.
- McMillan, J. H. (2001), *Essential assessment concepts for teachers and administrators*. Experts in assessment series, Guskey, T. R. & Marzano, R. J., Eds. Corwin Press: Thousand Oaks, CA.
- Sergiovanni, T. J. (2000). Standards and the lifeworld of leadership. *School Administrator*, 58 (8), 6-12.