

## The Numbers Game

By Deborah Jeffries Dusseau, David S. Hurst, and David Bitter

*Excerpted and adapted with permission from: Dusseau, Hurst, and Bitter, "The Numbers Game" Principal Leadership, September 2003, p. 49-54. Principal Leadership is a publication of the National Association of Secondary School Principals (NASSP). For more information concerning NASSP services and/or programs, please call (703) 860-0200, or visit <http://www.principals.org>. For the full article visit [http://www.ncacasi.org/documents/assess\\_game\\_prin\\_leader.pdf](http://www.ncacasi.org/documents/assess_game_prin_leader.pdf).*

### What Game Are You 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, you need to know what game you are playing - the cohort growth game or the program growth game?

	Cohort Growth Game	Program Growth Game
Definition	Compares the assessment results from a student group as it progresses from year to year	Compares the assessment results at the same grade level for multiple years
Example	Tracking 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	Evaluating the effectiveness of a new reading 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

Assessing program growth yields different information than cohort growth 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.

### What are the Rules of the Game?

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. While implementing this protocol, NCA CASI has identified some rules of the game that we call Green Light and Red Light Conditions.

<b>Green Light Conditions</b>	<b>Red Light Conditions</b>
Conditions that must be met before one can proceed	Conditions that should cause the users to "stop" and assess the influence of the occurrence on the results of the assessments
1. The <b>pretest (baseline)</b> and posttest instruments must be the same or psychometrically equivalent for any given assessment.	1. Beware of using tallies as measures. A tally has no fixed upper limit, making it difficult to determine the overall meaning of the figure.
2. Pretest and posttest assessments are conducted on the same grade level(s) of students.	2. Beware of using posttest data that has no true baseline.
3. Pretest and posttest assessments are <b>administered at the same time in the respective academic years.</b>	3. Beware of assessment results from non-random subsets of the population. While disaggregated data are useful in analyzing equity issues, they don't provide information about program growth for all students.
4. The pretest (baseline) assessment is administered close to the time that the implementation of the new programs or interventions	4. Beware of assessment results that come from a low or high ceiling assessment. Assessments that allow nearly all students to perform well or that sort out high achieving students yield scores that do not provide
5. The implementation period is sufficiently long for the new practices to have an authentic effect.	5. Beware of assessment results that are expressed as grade equivalents or stanines. They are difficult to manipulate statistically since they don't represent equal interval data.

## How Will 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.

**Use Multiple Assessments.** 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. 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.

**Convert Results to Standard Scores.** If a person accepts the premise that multiple measures are preferable over a single assessment, then the "game" gets complicated. If one assessment uses percentile scores as a metric and another uses raw scores, how can both assessment results be used and compared? 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 a 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.

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.

Even after data are converted to standard units, a major question remains: **Is the growth worth mentioning?** 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.

## 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.

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](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](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.