

**The National Longitudinal School-Level State Assessment Score Database -
NLSLSASD: Bringing it closer to the ideal***

**G. Gage Kingsbury
Northwest Evaluation Association**

December, 2005

Program evaluation is a very messy business in the best of times. The evaluation of federal programs and initiatives is even more difficult, because the scope of implementation is huge, and the desire to do what is right for students often overwhelms the need for evaluation. The National Longitudinal School-Level State Assessment Score Database (NLSLSASD) was developed to provide researchers access to information that might make the job of evaluating such large-scale programs manageable, although it will never be easy. The development of this database has clearly been a success. It has brought together information from a wide variety of sources into a common setting, and allowed researchers to use the data structure to guide and facilitate research, and to clarify language around that research. My purpose today is to discuss some of the strengths and weaknesses of the database, and to suggest some areas in which changes might be made to improve the evaluations that the database fosters.

Strengths of the NLSLSASD for Federal Program Evaluation

The NLSLSASD was designed in response to a need for a central repository of information for empirical research on the impact of a variety of federally-funded educational programs. This approach has a number of advantages, including the following:

Centrality: The primary advantage of the database is that it brings together a broad variety of information, imposes a clear data structure, and makes that information available in one place. In the past, researchers doing multi-state evaluation projects spent much of their time locating, collecting, and aligning information that the states had. While having the data in one place is helpful, having the data associated with a common data language is the crucial point. Knowing simple things like whether scores are scale scores or percentiles

Breadth: State assessment results include a broad cross-section of students in public education in the United States. With the advent of NCLB, states have begun to test all grades from 3 to 8 each year. NCLB also requires that virtually all students be tested, which expands the sample of

* Copyright © 2005 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.

Opinions and statements included in the paper are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Planning Committee for the Symposium on the Use of School-Level Data in Evaluating Federal Education Programs or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

students even more. This amazing breadth enables strong comparisons of performance of subgroups of students within states and across years.

State-Level Consistency: Since NCLB has caused states to administer the same type of test to all of their students (with some exceptions) the database provides comparable information on student achievement across grades, within each state. While not all states have vertical scales associated with their assessments, just having consistent types of data is a major step forward. In other words, researchers don't have to try to connect the dots from percentile ranks in one grade to unrelated scale scores in another grade.

Motivation Consistency: In the past, researchers have struggled with the need to add a measure of student achievement to their evaluations. If this measure wasn't perceived as useful by the students and teachers, the differential motivation of students taking the assessments caused additional noise to be added to the study. The use of state assessments in NLSLSASD may not assure motivation of all students, but should serve to homogenize the amount of motivation for students within a state, particularly following NCLB passage.

Longitudinal Information: Most evaluation studies suffer from an absence of information prior to the start of the study, and from a need to collect information during the first year of implementation of a program. NLSLSASD provides information that often precedes the adoption of a particular program, or is available during a change in a program. This makes the information in the database particularly useful for time-series research (sometimes mislabeled quasi-experimental research).

Weaknesses of the NLSLSASD

While the NLSLSASD has outstanding strengths, any database is limited by the data available and by the purpose for which it is designed. The limitations of the NLSLSASD to enable program evaluation of federal programs include the following:

“Quasi-Experimental”: The database contains a wealth of information that was collected without random assignment of students (or schools, or districts, or states) to treatment categories. Longitudinal designs are much better suited for the evaluation of large-scale programs than the medical/agricultural model. The medical model of random assignment and double-blind trials doesn't really work in most educational settings since teachers need to know what they are teaching. This suggests a need to clarify what constitutes a powerful approach to identifying the impact of an educational program. That said, random assignment is the model that is preferred by the Clearinghouse. It would be useful to discuss changes in the process for identifying high-quality studies with the Clearinghouse.

Implementation: In any evaluation study, knowledge of the degree to which a program is implemented in different settings is imperative. This implementation level becomes one of the formative measures in the study, and should be used to inform the summative measures. In the research done to date with the NLSLSASD, this information has been captured after the fact. It

would be preferable to require a survey of implementation to accompany the initial implementation of a program and changes in a program as they occur.

Inconsistency before and after NCLB: The longitudinal data from state to state tends to be somewhat inconsistent before and after NCLB went into place. While some states maintained the same testing system after NCLB, many modified their testing system or began using new test providers. At the same time, teacher and student motivation toward the results of the tests may have changed as a result of the new legislation. As a result, results from state assessments might be particularly susceptible to change since NCLB went into effect.

Different tests: From one state to another, different tests are used, and different proficiency levels are used as the benchmark for student performance. This means that comparisons across states are weakened by the differences in the tests and test designs.

Different score types: The NLSLSASD has the nice feature of identifying the types of scores that are available for each test. While this is quite useful, it may lead to problems with comparisons. For example, percentile ranks are not appropriate for many statistical analyses and aren't comparable across tests. Proficiency levels differ in meaning from state to state and as a result the percentage of students reaching a level of proficiency isn't comparable. As a result, the analyses that are facilitated by the database should be undertaken with extreme caution to avoid misinterpretation of the data.

Differential accuracy: Tests differ substantially in the degree of accuracy with which they measure students at different achievement levels. For instance, a fixed-form test that is very accurate for high-achieving students may be very inaccurate for low performing students. While this is a well known phenomenon, it is fairly consistently ignored by evaluators. As long as the evaluator chooses a single instrument that it is appropriate to measure change in student performance, this usually isn't a big problem. However, when multiple instruments are being used and those instruments vary in their measurement characteristics, unexpected biases can occur in any study.

Suggestion for improvements to NLSLSASD

Score validity indicator: One element that might be considered for addition is measure of the average standard error for a test score at each level of aggregation. This would enable the use of statistical procedures that allow for errors of measurement. A second element that might be even more useful is an indication of the percentage of invalid scores for test scores at each level of aggregation. Invalid scores (those from students with near-perfect or near-chance scores) can add substantial bias to statistical measures (positive bias near chance, and negative bias near perfect). In most analyses, these scores are treated as usable, and may result in over or under-estimates of effect sizes.

Example. An evaluation using school-level average scale score changes from one grade to the next from a single state might indicate a negligible effect size for a particular math program being used with students struggling with mathematics. If the pretest has a

substantial percentage of invalid scores with students scoring near chance, it is likely that pretest achievement was lower than the test scores indicated. This may have reduced the observed effect size, falsely indicating that the program wasn't effective.

Implementation survey: To enable researchers to make better use of achievement data to investigate the impact of a program, it would be very useful to add an indicator of implementation to the database for each major program that might be of interest, for each year of implementation. Without such an indicator, the treatment in the evaluation (the educational program, in this case) will not be consistent across settings. This may reduce the overall observed impact of a program, but it also reduces our ability to evaluate the ability or interest in implementing the program.

Example. In a large-scale program designed to provide summer-school reading classes for students in grades 2 and 3, schools are given small grants, books, and lesson plans. The overall analysis of change over three years might indicate a positive change in student's average reading achievement. This is useful information, but consider how much more useful it would be to know that some schools put the books in the library, ignored the lesson plans, and made little improvement in reading. At the same time, other schools took full advantage of the program and made significantly greater gains. The information is useful to clarify the analysis, but more useful to suggest a need for more precise implementation procedures.

Avoid cutoff score analysis: From a psychometric point of view, one of the worst types of evaluation that can be done is to count the number of students above or below a cutoff score. Any well-made test is designed to identify a student's achievement as accurately as possible. The use of a cut-off score throws away almost all of the information that is gathered by the test. It also reduces our ability to differentiate fine gradations of performance in our students, and as a result reduces our ability to examine the impact of educational programs. Identifying whether students meet or exceed a proficiency cutoff score can be very useful for several types of categorical decisions such as whether or not a student graduates or is retained in the current grade. In professional settings, this type of identification is also quite useful for certification and licensure decisions. For evaluation, though, the percentage of students crossing a hurdle is at best a very gross measure, and is hardly ever the best choice for a program evaluation. This isn't a suggestion for a change in the database, but it is a suggestion for change in advised analyses.

Example. Consider a multi-state evaluation of a high school program to enhance science learning in gifted students using the enquiry model. The results indicate little change in the percentage of students identified as proficient in science using state-defined cutoff scores. This information isn't very useful unless we are sure that the proficiency cutoff score was close to the performance level at which we expected the program to have an effect. Using scale scores, or raw scores, or even percentile ranks would allow us to identify the impact of the program more precisely in almost every evaluation.

Test change indicator: The impact of NCLB has been wide-ranging within state assessments, and one of the small consequences is that states have re-evaluated their assessment programs and changed test vendors as they filled out the grades and standards to be assessed. There is a far amount of research that indicates increases in test scores during the first years of use, regardless of

changes in other programs or materials. Capturing the changes in both the tests used and the content areas assessed in each state would enable more precise evaluations and avoid mistaking test impact for program impact.

Example. In a multi-year study of student achievement in school-wide Title I programs in a single state, the results may indicate a drop in performance from the first year of the study to the second in all groups followed by steady increase in subsequent years. If we knew that the state had changed its content standards and implemented the change in its assessment system during the second year of the study, this overall pattern might be partially attributable to test change. If this were the case, the discussion of the evaluation might take a substantially different tone and provide more enlightenment to the readers.

Connections to other data sources: The lowest unit of analysis available within the database is the school and this is quite appropriate for many evaluations. Imagine how the scope of the database could be extended through connections with other databases that might include class and even student level data. Since many large-scale programs are not designed to be implemented school wide and may not be designed to affect the achievement of all students, identification of anonymized teacher and student level information could enable researchers to isolate program effectiveness much more precisely. While most other databases are much more limited, the addition of some classroom data could still improve analysis substantially.

Example. The NWEA Growth Research Database has over 30,000,000 assessment records of student longitudinal performance, some dating back to 1996. These data are related to teacher and school information and each school is related to the NCES school identifier and through this to all the common core data. All data are anonymized and restricted where use could identify any student. If this database were related to the NLSLSASD it could provide researchers with a subset of their analysis sample that could be used to pinpoint impact of programs in particular groups of students within certain settings. While databases of this type have non-representative samples, the subgroup analyses done after the primary NLSLSASD analysis could be quite powerful.