

Discussant Remarks

Laurie Wise
HumRRO
December 9, 2005

The central question of this symposium was “How can the SSASD be used to evaluate federal education programs?” A number of different ways were identified by presenters, including:

- Using school-level information to build a sampling frame for identifying matched samples of schools for randomized trials or other program effectiveness studies.
- Providing historical, pre-implementation data on school characteristics and test scores prior to implementation of specific programs or policies.
- Analyzing hypotheses about causal inferences following the autopsy model for causal determination suggested by Michael Scriven.
- Notwithstanding concerns expressed by Bob Linn, David Thissen, and others about the comparability of scores across states and over time, school means can provide useful outcome information for program evaluations in some circumstances. Don’t let “perfect” be the enemy of “good enough!” An “approximate” answer is still useful and many analyses may focus on within-state, within-year comparisons.

While a number of important uses of the school-means data base were identified, so too were a number of limitations. These limitations must be recognized in current uses and steps to minimize their impact should be taken to improve the effectiveness of the school-means data base in evaluating federal programs. Issues identified by the presenters that warrant further exploration include:

- The validity of test scores. State assessment scores may not be a perfectly valid measure of the outcome(s) that the federal program being evaluated was designed to achieve. The methods of linking scores over time and across states often show different relationships for different groups of students (lack of population invariance). Consistent relationships are particularly critical in investigating efforts to reduce achievement gaps among different demographic groups.
- Below-school-level programs. School-level means are limited as outcome indicators for programs that are targeted to specific populations or classes within the school.
- Not (yet) possible to assess individual student growth. Using year-to-year and grade-to-grade comparisons of school means to assess student growth is limited because of the mobility of individual students and also drop-out and retention policies. The grants program to help states build longitudinal data

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.

systems described by Kashka Kubzdela could lead to improved possibilities for longitudinal student data.

- Timeliness. It is not clear how soon after state assessments are conducted data files will be available. Current plans for assembling the data are ambitious and it could take a considerable time to meet targeted standards for inclusion. School mean data that are several years old would limit their usefulness in evaluation more recent federal programs. The school-mean data assembled in conjunction with NAEP, while not covering every state, are already available for recent assessment years.
- Between-state comparisons and the impact of state policy on implementation and effectiveness of federal programs. The effectiveness of the federal programs being evaluated may vary from state to state as a function of difference in state policies and practices. While several presenters expressed concern about the impact of differences among states in test content, data presented by Don McLaughlin suggests that this may be less of a problem for school-level data than it would for individual student data. School means in reading were highly predictive (correlation of .86) of school mathematics means. Correlations across different measures of mathematics are likely to be even higher.

The presenters also suggested a number of possible enhancements to improve the usefulness of the database. Additional types of information that might be particularly useful include:

- Data on characteristics of the state assessments including dates when a new assessment is introduced, test contents and reliability information, and possibly data on the alignment of test content to the content of the NAEP assessments
- Data on program characteristics and implementation information, including both the federal programs that might be evaluated and state programs that could interact with these programs
- Standardization of test scores over time and across states, converting from arbitrary metrics to effect sizes and possibly using NAEP results to adjust for state differences and differences over time in proficiency standards
- Other outcomes, such as educational attainment and dropout rates, participation in and scores on Advanced Placement tests.

Finally, if efforts to develop and maintain the database are to succeed, the Department must demonstrate advantages for states? If supplying data for this endeavor simply becomes another federal requirement, cooperation from the states may be limited and quality and timeliness could suffer. If states see advantages from opportunities for comparisons to other states, they are more likely to become a willing partner.

The Department also needs to decide whether a completely centralized effort is the best way to ensure timeliness and quality. An alternative would be establish a more distributed system creating multiple, linkable data bases that could each be developed on their own timeline.