

## A Proposal to Enrich the School-Level Data<sup>\*</sup>

Donald B. Rubin  
Harvard University  
rubin @ stat.harvard.edu

Many important issues have been discussed thus far. I will focus on one that has not received as much attention as it might have. That is, how do we create an improved data source using the data that are currently available as the basis? To do this, I will draw on methods that have been developed over the past three decades to handle missing and incomplete data, including work in Holland and Rubin (1982) on Test-Equating. My approach could be called an “active” one because I propose that we should collect some new data, in contrast to the current discussions, which could be called “passive” in that they deal entirely with how to use the data we already have. Thus, I propose considering an approach for enriching the current data sets.

The conceptual starting point for this brief discussion is: What are “all” the data that we wish we had from all the state-level assessments? I believe that the “ideal” data set, more than we could ever hope to have, can be described in the following way. Let all the kids in all the states form the rows of a giant “complete-data” matrix, where the kids from

---

<sup>\*</sup> 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.

Discussant Remarks – Professor Donald B. Rubin  
NRC Symposium on Use of School-Level Data to Evaluate Federal Education Programs

the same states are adjacent, thereby creating 50 groups of rows, in alphabetical order by state, for example. Each states' test items form the columns of this giant data matrix, where the items from each state are adjacent, thereby creating 50 groups of columns. The complete-data matrix is thus a 50 by 50 matrix of blocks, where within each of the 50 row groups are distinct rows for individual kids, perhaps grouped further by school, and within each of the 50 column groups are distinct columns for individual test items, perhaps grouped further by sections of each test. The entry in each cell of this giant complete-data matrix gives the score that kid would make on that item if that kid could took that state test first, that is unpracticed by taking any of the other states' tests. Thus, this complete-data matrix is totally hypothetical, and could never be directly observed, although any entry of it could be. Currently, each state provides data on some sample of kids in its state but only on its test: state  $i$  provides data only on test  $Y_i$ , and maybe some other test such as NAEP. As a result, the observed data on the  $Y$ s essentially form a block diagonal matrix. Each state produces data on its test for its kids, unpracticed by tests from the other states.

In fact, however, this entire complete-data matrix can be estimated from observable data, although only very poorly from the data that we currently have observed, because at present, no kids take any tests from states other than their own. But if we did have the giant complete-data set as just described, we could compare students and states on any subset of test items from any of the states tests, even with respect to hypothetical tests that were never in fact given to anyone. For an example, a 50 section test created by

selecting one section from each state's test could be used to compare performance of the kids. Therefore, estimating this complete-data matrix must, in some sense, be our actual immediate objective, and how to do this is the thrust of this brief discussion. Having data on covariates, like NAEP, that are currently at least partially observed in every state, can be very helpful for estimating the joint distribution of the  $Y$ s (e.g., see Rubin and Thayer, 1978), but that is not the focus of this proposal.

The crucial idea is to have some kids from each state take, in addition to their own states' test, some portion of one or two other states' tests. The basic idea is described in Holland and Rubin (1982) and is called "section pre-equating". It can be done in such a way so as to account, under certain assumptions, for practice effects, by imbedding the other states' sections in their own states' test for a sample of kids. The details about the actual design are entirely beyond the scope of this brief discussion, but are discussed in a variety of places, including Holland and Wightman (1984) in Holland and Rubin (1982), Raghunathan and Grizzle (1995), Raessler (2002), etc.

As long as each test item (or section) from one state appears with each test item (or section) from another state on a reasonable-sized subsample of kids in each state, the within-state means, variances, and correlations can be uniquely estimated, and under an assumption of joint normality for all sections of the tests, the entire distribution can be estimated. Normality is often a reasonable assumption when test scores are averages across many items. In fact, more general ellipsoidally symmetric distributions can be fit

with some more computational effort (e.g., see Liu and Rubin, 1998). The missing tests scores for each states' sections would then be multiply imputed (Rubin 1987, 2004), just as “missing” test scores are multiply-imputed in NAEP.

From this multiply-imputed ideal data set, all inferences that could be drawn from the ideal complete-data set described earlier can be drawn. Of course, there will be additional uncertainty with the multiply-imputed data set relative to the ideal data set because of the missing data that are created by the sampling scheme used to collect the data on out-of-state test sections, but this extra variance is reflected by the variance across the multiply-imputed data sets. And of course, the inferences will be somewhat reliant on the normal or ellipsoidal models used to create the multiple imputations, but the same caveats apply, for example, to NAEP itself, and a tremendous amount of statistical evidence supports the propriety of the approach (e.g., see Rubin, 1996).

The payoff, if this approach is used, is the ability to address very complex questions with a wonderfully rich data base. The expense to create this multiply-imputed data set is nearly all in designing the extra tests and giving them to the kids. But perhaps the benefits could be worth the extra expense of data collection. I hope so, but I am in no position to assess these trade offs, and so I leave those decisions to others. In any case, I hope that this brief presentation of ideas provides food for thought for those more knowledgeable about implementation issues than I am.

## REFERENCES

Discussant Remarks – Professor Donald B. Rubin  
NRC Symposium on Use of School-Level Data to Evaluate Federal Education Programs

Holland, P.W. and Rubin, D.B. (1982). *Test Equating*. New York: Academic Press, Inc.

Holland, P.W. and Wightman, L.E. (1984). “Pre-Equating: A Preliminary Investigation.” In *Test Equating*. (P.W. Holland and D.B. Rubin, eds.). New York: Academic Press, Inc, 271-306 (with discussion).

Liu, C.H. and Rubin, D.B. (1998). “Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data.” *Biometrika* **85**, 673-688..

Raessler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. New York: Springer Verlag.

Raghunathan, T.E. and Grizzle, J.E. (1995). “A Split Questionnaire Survey Design.” *Journal of the American Statistical Association* **90**, 55-63.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Rubin, D.B. (1996). “Multiple Imputation After 18+ Years.” *Journal of the American Statistical Association* **91**, 473-489, with discussion, 507-515, and rejoinder, 515-517.

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Reprinted with new appendices as a “Wiley Classic.” New York: John Wiley and Sons.

Rubin, D.B. and Thayer, D.T. (1978) “Relating Tests Given to Different Samples.” *Psychometrika* **43**, 3-10.