Data Adjustment for Educational Assessment

Jiahe Qian, Educational Testing Service
Rosedale Road, MS 02-T, Princeton, NJ 08541

KeyWords: Educational survey, Iterative Proportional Fitting Procedure, Post-stratification, Equating

When non-equivalence between two tests is found in comparison of test scores or merging test data from two samples, data adjustment is always necessary. In assessment field, adjustment of test scores, equating, is to align two sets of scores from two tests on a common scale. Non-equivalence could be generated by the discrepancies in score distributions in the corresponding aggregates between two assessments. These aggregates are formed by multiple variables, such as areas, school type, etc. If trends of the scores of aggregates are of interest, data adjustment for merging data should be based on aggregates instead of the total.

Some practices successfully applied data merged from different frames. In the study of Aptitude Score Distribution, Spencer et al. (1991) used data from High School and Beyond and National Longitudinal Study to improve the estimation of Armed Forces Qualifying Test Score Distributions in Counties and Battalion Regions. Some studies (Boruch, 1999) have been done to link test scores from available standardized tests to be compared to each other and the National Assessment of Educational Progress (NAEP).

This article will mainly discuss data adjustment in merging of educational assessments. The results will be based on NAEP data. The NAEP assessment programs usually collect samples separately for state assessments and national assessments, though the tests of same subject for students of same age are usually identical. One of the interested research topics is how to link the data from two programs together in NAEP assessments.

One case of non-equivalence

One example of such discrepancies is from the NAEP 1998 Reading Assessment. The program collected samples separately for State and National assessments, yet students were of same age and took same tests. The State sample of public schools consisted of grade-eligible public school students from 39 jurisdictions participated. Table 1 lists nine states with significant differences between two tests in mean scales or the relative difference are larger than 5 percent. This example show that linking of two tests will not ensure the equivalence of test scores for aggregates within the population although the State scores had been equated to the national scores (Allen, 1998). This example show linking of two tests will not ensure the equivalence of test scores for aggregates within the population.

Several factors could cause these discrepancies. First, the two data sets are selected by different sampling schemes. In addition, the test conditions, student motivation to perform, and school participation are usually different in administrations of assessments. Moreover, Mislevy (1998) pointed out that the IRT models for state and national samples could vary when what is being conditioned on vary. Of course, neglect of discrepancies would introduce intolerable bias in estimation. Based on IRT theory, Holland (1998) studied the factors that cause two tests not parallel.

Methodology of data adjustment

To obtain two relative equivalent data sets, two of the basic issues in data adjustment in merging two assessments are: 1) the adjustment of sampling weights with fixed marginals and, 2) the adjustment of test scores for population and corresponding aggregates. The data adjustment to test scores, which brings one assessment onto the same scale of the host assessment, is also called equating. The host assessment is the one with relatively standard test conditions.

1) The adjustment of sampling weights with fixed marginals

Let $\{n_i^A\}$ and $\{n_i^B\}$ are the marginal distributions of a two by two table and, $\{Z_{ij}\}$ are the inclusion probabilities of m_{ij}. The adjustment of sampling weights is to reduce the possible bias of estimation. The Deming-Stephan approach (1940) is to minimize

$$\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}(\hat{m}_{ij} - m_{ij})^2 / m_{ij},$$

subject to the marginal totals, $m_{ij} \geq 0$. The marginal distributions of variables are demographic variables but not dependent variables in study. The main demographic variables involved in NAEP study are region, gender and ethnicity.

The well-known method to obtain the approximate

solution is the Iterative Proportional Fitting Procedure (IPFP) that was introduced by Deming (1940). The $\hat{m}_{ij}$ obtained by Deming-Stephan algorithm are also the approximate solutions of the maximum likelihood equations satisfying marginal conditions (Haberman, 1978).

The variety of statistical issues of IPFP, including convergence of the IPFP-algorithm in the finite discrete case, were included in papers by Brown (1959), Bishop and Fienberg (1969), Ireland and Kullback (1968), Fienberg (1970) and Csiszar (1975). Ruschendorf (1993) found a general convergence proof of the IPFP-algorithm. Recently, the algorithm was modified under different assumptions (Ruschendorf, 1996).

To check the effects of adjustment of weights, we compare achievement level scores before and after adjustment. The achievement level scale is designed to categorize student achievement within ranges of Basic, Proficient and Advanced. Then reporting percentages of students attaining specific NAEP achievement levels provides useful information. In Table 2 lists the results of percentage of students at or above Basic achievement level after adjustment of weights. Some discrepancies of the percentages between two samples clearly exist. This suggests that further adjustment of test scores is necessary.

2) *The adjustment of test scores for corresponding aggregates*

The traditional role of data adjustment of test scores, equating, is to bring one assessment onto the same scale of the host assessment with same means and variances as whole (Lord, 1980). Several equating strategies are explored: a) applying the equating approach to the aggregates which are formed by sampling frames or post-stratification, and, b) applying percentile equating or linear equating method subject to fixed marginal distributions of main demographic variables.

a) Equate aggregates formed by post-stratification

Post-stratification equating is based on two sets of scale score distributions for respective aggregates. As in NAEP, linear equating is applied in this analysis. For details, see Allen (1998). To link the State and national scales of the 1998 NAEP Reading Assessment, one set of scores, in the scale of State Assessment, is obtained from the State sample of the aggregate. The other is based on the subsample of the aggregate, if it exists, from the national sample in the reporting National scale. For each aggregate, the State Assessment and national scale scores were made comparable by constraining the mean and standard deviation of the two sets of estimates to be equal.

To check the effects of equating, we also compare achievement level scores before and after adjustment. Although linear equating ensures equality of means and standard deviations for two samples, it will not assure same shapes of distributions of estimated scores from the two assessments. Since two samples are from same target population, to justify strong claims of comparability for the state and national scales, the distributions of estimated scores based on two samples should be similar in shape. Also in Table 2, we can find the results of percentage of students at or above Basic achievement level after applying linear equating aggregates formed by post-stratification. Although some discrepancies of the percentages between two samples still exist, the sizes of the discrepancies are reduced.

b) Equate aggregates subject to fixed marginal constrains

Instead of post-stratifying the population, we can apply equating aggregates subject to fixed marginal distributions of main demographic variables. The issue can be treated as the issue of a restricted least squares estimators (Dykstra, 1985). It is a generalization of the Deming-Stephan approach of adjustment. Adapted iterative proportional fitting procedure for equating is used to obtain the solutions of restricted least square problems. The bridge between the two issues is that the IPFP algorithm is an analogue to the alternation algorithm that was introduced in the case of Hilbert space by von Neumann (1950) and Aronszajn (1950). The problem of restricted least squares estimators is to minimize

$$\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}(\hat{y}_{ij} - y_{ij})^2 w_{ij},$$

subject to the marginal conditions

$$K_i = \{\hat{y}_i ; f(\hat{y}_i, y_i) = 0\}$$

where i = 1, 2,... I refer to the marginal variables.

The procedure converges for linear equating approach and Deming-Stephan approach (Ruschendorf, 1996). However, for linear equating, the marginal conditions become

$$K_i = \{\hat{y}_{ij} ; \hat{y}_{ij} = A_i \cdot y_{ij} + B_i, j=1, 2, ..., J\}.$$

For Deming-Stephan approach, y_{ij} represents weight m_{ij}. The marginal conditions become

$$K_i = \{\hat{m}_{ij} ; \hat{m}_{ij} = \sum \hat{m}_{ijk} = m_{ij} \wedge m_{ij} \geq 0, j=1, 2, ..., J\},$$

where j is the index of category for marginal variable i, and k is the index of cases in cell ij.

After applying equating aggregates subject to fixed marginal constrains, in Table 2, the discrepancies of the percentage of students at or above Basic achievement level are decreased. The effects are similar to results of linear equating aggregates formed by post-stratification.

Conclusions

The empirical data in NAEP show that non-equivalence between two samples of the same type surveys. Data adjustment becomes necessary when non-equivalence is found, which could be caused by discrepancies in distributions in the corresponding aggregates between two samples. To obtain two relative equivalent data sets, two levels of data adjustment can be applied. First, to reduce bias in estimation, adjust sampling weights by Deming-Stephan algorithm. Second, adjust test scores at corresponding aggregates by applying linear equating aggregates formed by post-stratification or applying equating aggregates subject to fixed marginal constrains.

The results based on the 1998 Reading Assessments show that the equating has worked satisfactorily.

References

Allen, N. et al. (1999). *The 1998 NAEP Technical Report.* Washington DC: National Center for Education Statistics (in Press).

Aronszajn, N., (1950). Theory of Reproducing Kernels. Trans. Amer. Math. Soc. 68, 377-404.

Bishop, Y.M., and Fienberg, S.E., (1969). Incomplete two-dimensional contingency tables. Biometrics 55, 119-128

Boruch, R. & Terhanian, G., (1999). Putting studies , Surveys, and Data Sets Together: Linking NCES Surveys to One Another and to Data Sets from Other Sources. Presentation at the 1999 American Educational Research Association annual meeting, Montreal, Canada.

Brown, D.T., (1959). A note on approximations to discrete probability distributions. Inform. and Control 2,386-392

Csiszar, I., (1975). I-divergence geometry of probability distributions and minimization problems. Ann. Probab. 3, 146-158.

Deming, W.E., & Stephan, F.F., (1940). On A Least Squares of Adjustment of A Sampled Frequency Table When the Expected Totals Are Known. Ann. Math. Statist. 11, 427-444.

Dykstra, R.L., (1983). An Algorithm for Restricted Least Squares Regression. Journal of American Statistical Association. 78, 837-842.

Fienberg, S.E., (1970). An Iterative procedure for Estimation in Contingency Tables. Ann. Math. Statist. 41, 907-17.

Fuller, W.A., and Burmeister, L.F., (1972). *Estimators for Samples Selected for Two Overlapping Frames*, Proceedings of the Social Statistics Section, American Statistical Association, 245-249.

Haberman, S., (1978). Analysis of Quantitative Data. Vol. 1: Introductory Topics. New York: Academic.

Holland, P. et al. (1998). *Uncommon Measures: Equivalence and Linkage of Educational Tests.* Washington DC: National Academy of Science.

Hartley, H.O., (1974). *Multiple Frame Methodology and Selected Applications*, Sankhya, 99-118.

Ireland, C.T., and Kullback, S. (1968). Contingency tables with given marginals. Biometricka 5, 179-188.

Johnson, E., Mazzeo, J., and Carlson J. (1995). Sampling Issues for 1996. Distributed at February 9, 1995 Meeting on NAEP Sampling Issues, Washington, D.C.

Lord, F. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: LEA.

Mazzeo, J. and Williams ,P. (1995). Summary of 1996 Sampling Issues Meeting. Memorandum for Gary Phillips, 1/17/95.

Mislevy, R., (1998). A note on merging national and state samples. Princeton, NJ: Educational Testing Service.

Ruschendorf, L., (1993). Convergence of the iterative proportional fitting procedure. Ann.Statist.

Ruschendorf, L., (1996). Development on Frechet-Bounds, Distributions with Fixed Marginals and Related Topics IMS Lecture Notes- Monograph Series Vol. 28, 1996

Spencer, B.D. (1996). School and Student Sampling in the 1994 TSA: An Evaluation. Stanford, CA: The National Academy of Education, in press

Spencer, B., Nordmoe, E., Qian, J.,and Haberman, S., (1991). *Aptitude Score Distribution Study--Phase 1: Sampling Merging and Direct Estimation of Armed Forces Qualifying Test Score Distributions in Counties and Battalion Regions*, Research Report, NORC.

von Neumann, J., (1950). Functional Operators, Vol.

Table 1. Weighted mean scale scores of states
for the 1998 Reading Assessments, Grade 4 Public schools

	State Assessment (N=89,164)			National Assessment (N=6,300)		
	Weighted percentage	Mean scale scores	SE of mean scale scores	Weighted percentage	Mean scale scores	SE of mean scale scores
Arkansas	1.2	209	1.5	1.1	224.0	***
Delaware	0.3	212	1.3	1.0	224.0	2.2
Hawaii	0.5	200	1.8	0.9	193.0	2.0
Maine	0.6	225	1.2	0.5	219.6	***
Montana	0.4	226	1.7	1.2	215.2	***
New Hampshire	0.6	226	1.3	2.4	238.9	1.7
New Mexico	0.8	206	2.0	0.8	230.9	3.3
Tennessee	2.6	212	1.5	1.5	221.3	***
Wisconsin	2.2	224	1.1	1.9	209.9	10.2
Virgin Island	0.1	178	1.9	NA	***	***

Table 2. Percentage of students at or above Basic achievement level
with different adjustments for some states
in the 1998 Reading Assessments, Grade 4 Public schools

	State sample	State + National samples			
	Original data	No adjustment	Adjustment of weights	Adjustment of scale score (post-stratification)	Adjustment of scale score (IPFP)
Arkansas	55.0	55.7	55.4	55.4	55.3
Delaware	57.0	58.5	57.9	57.8	57.6
Hawaii	45.0	44.5	44.7	44.7	44.7
Maine	73.0	72.7	72.8	72.8	72.8
Montana	73.0	71.5	72.0	71.9	71.9
New Hampshire	75.0	77.1	76.4	76.2	76.2
New Mexico	52.0	53.3	52.7	52.3	52.2
Tennessee	58.0	58.3	58.2	58.1	58.2
Wisconsin	72.0	71.2	71.6	71.6	71.5