Listening. Learning. Leading.®

Research Report
ETS RR–13-11

Poststratification Equating Based on True Anchor Scores and Its Relationship to Levine Observed Score Equating

Haiwen (Henry) Chen

Samuel A. Livingston

May 2013

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

Poststratification Equating Based on True Anchor Scores and Its Relationship to Levine Observed Score Equating

Haiwen (Henry) Chen and Samuel A. Livingston

ETS, Princeton, New Jersey

May 2013

Action Editor: Matthias von Davier

Reviewers: Shelby J. Haberman

Abstract

This paper presents a new equating method for the nonequivalent groups with anchor test design: poststratification equating based on true anchor scores. The linear version of this method is shown to be equivalent, under certain conditions, to Levine observed score equating, in the same way that the linear version of poststratification equating is equivalent to Tucker equating. Some issues related to this result are discussed.

Key words: nonequivalent groups with anchor test design, poststratification equating based on true anchor scores, Levine observed score equating

The nonequivalent groups with anchor test (NEAT) design is commonly used for equating the scores on different test forms in many high-volume, high-stakes testing programs. Choosing an equating method is often a major concern for the statisticians who work on these programs. Among the equating methods that do not require the strong assumptions of item response theory, the most commonly used are these:[1]

- Chained linear equating
- Chained equipercentile equating
- Tucker equating
- Levine observed-score equating
- Levine true-score equating
- Poststratification equipercentile equating (PSEE), also called *frequency estimation equipercentile equating*

Among these, only the chained equipercentile method and the poststratification equipercentile method allow for the possibility of a nonlinear equating relationship. This limitation of the other methods has important practical consequences, because when test forms differ in difficulty, the equating relationship is often curvilinear.

For several years, psychometricians have been attempting to develop a curvilinear analogue to the Levine method. These attempts include

- a method in which the equated score determined by PSEE is modified by adding the difference between the equated scores determined by two linear equating methods: the Levine observed-score method and the Tucker method (von Davier, Fournier-Zajac, & Holland, 2006),
- two methods in which PSEE is modified by transforming the score distributions with the mean-preserving linear transformation (Chen & Holland, 2009; Wang & Brennan, 2007), and
- chained true-score equipercentile equating, a chained equipercentile equating of estimated true-score distributions (Chen & Holland, 2008).

The first three of the four methods described above use some kind of mathematical manipulation, either on the score distributions or on the equating function. They are not analogous to the Levine method in the way that PSEE is analogous to the Tucker method, producing the same equating function when the assumptions of Tucker equating are met (Braun & Holland, 1982).

This paper defines a new equating method—*poststratification equipercentile equating based on true anchor scores* (PSEE$_{\text{TA}}$)—and demonstrates that it has the following properties:

1. When the assumptions of Levine equating are met, the linear equating from the cumulative distributions produced by PSEE$_{\text{TA}}$ is Levine observed score equating.
2. By applying a mean-preserving linear transformation to the joint distribution of observed test scores and anchor scores, one can estimate the joint distribution of observed test scores and true anchor scores. The PSEE of these modified distributions will then approximate the PSEE$_{\text{TA}}$. The computation is identical to that of the method previously referred to as curvilinear Levine observed score equating (Chen & Holland, 2009).

Equating Methods

In this paper, *X* and *Y* will represent scores on the test forms to be equated, with population ***P*** taking Form X and population ***Q*** taking Form Y, while A will represent the score on an anchor test taken by both populations ***P*** and Q. The equating relationship between *X* and *Y* is to be determined for the synthetic population ***S***, defined as a weighted mixture of populations ***P*** and ***Q***, represented in the proportions *w*, for population ***P***, and $(1-w)$, for population ***Q***. The symbols T_X and E_X will represent the true-score and error components of score *X*, with similar notation for score *Y* and anchor score *A*.

The general form of a linear equating from *X* to *Y* in population ***S*** is the following:

$$y = \mu_S(Y) + \frac{\sigma_S(Y)}{\sigma_S(X)}\left[x - \mu_S(X)\right]. \tag{1}$$

Tucker Equating

Tucker equating consists of Equation 1 with estimates of the means and standard deviations that are based on the following assumptions (see Kolen & Brennan, 2004, p. 106):

1. The regression of X on A and the regression of Y on A are population invariant. That is, they are the same in population $\boldsymbol{S}$, where they cannot be directly observed, as in populations $\boldsymbol{P}$ and $\boldsymbol{Q}$, where they can be observed.
2. The conditional variance of X given A and the conditional variance of Y given A are population-invariant.

Levine Observed Score Equating

Levine observed score equating also has the form of Equation 1, but with estimates of the means and standard deviations that are based on different assumptions (see Kolen & Brennan, 2004, p. 110):

1. The regression of T_X on T_A and the regression of T_Y on T_A are population-invariant.
2. The variances of the error components of X and of Y are population-invariant.
3. True scores on the anchor and on the tests to be equated are perfectly correlated.
4. The variance of the error component of A is population-invariant.

Generalized Levine Observed Score Equating

Generalized Levine observed score equating is defined as a linear equating of X to Y in population $\boldsymbol{S}$, based on Equation 1, with estimates of the means and standard deviations that are based on the following assumptions:

- Assumption 1—The regression of X on T_A and the regression of Y on T_A are population-invariant.
- Assumption 2—The conditional variance of X given T_A and the conditional variance of Y given T_A are population-invariant.

Under the usual classical test theory assumption that the errors in X and A are uncorrelated with their true scores and with each other, the slope of the regression of X on T_A is

$$\frac{\mathrm{cov}(X, T_A)}{\mathrm{var}(T_A)} = \frac{\mathrm{cov}(T_X, T_A) + \mathrm{cov}(E_X, T_A)}{\mathrm{var}(T_A)} = \frac{\mathrm{cov}(T_X, T_A)}{\mathrm{var}(T_A)}, \qquad (2)$$

and the conditional variance of X given T_A is

$$\operatorname{var}(X \mid T_A) = \operatorname{var}(X)\left[1-\rho^2(X,T_A)\right] = \operatorname{var}(X)\left[1-\frac{\operatorname{cov}^2(X,T_A)}{\operatorname{var}(X)\operatorname{var}(T_A)}\right] = \operatorname{var}(X) - \frac{\operatorname{cov}^2(X,T_A)}{\operatorname{var}(T_A)}$$

$$= \operatorname{var}(X) - \frac{\operatorname{cov}^2(T_X,T_A)}{\operatorname{var}(T_A)} = \operatorname{var}(X) - \rho^2(T_X,T_A)\operatorname{var}(T_X). \tag{3}$$

The same results occur for *Y*.

Levine observed score equating then becomes a special case of generalized Levine observed score equating, with two additional assumptions:

- Assumption 3—Both T_X and T_Y are correlated perfectly with T_A.
- Assumption 4—The variance of the error component of *A* is population-invariant.

Assumptions 3 and 4 are the assumptions identified above as Assumptions 3 and 4 of Levine equating.

Assumption 1 of Levine equating is identical to Assumption 1 of generalized Levine equating, because the linear regression of T_X on T_A is the same as the linear regression of *X* on T_A. (See Equation 2.)

Assumption 2 of Levine equating says that the error variances of test scores *X* and *Y* are population independent. If Assumption 3 is true, then this assumption follows from Assumption 2 of generalized Levine equating. Note that if $\rho(T_X, T_A) = 1$, then, from Equation 3, the conditional variance of *X* given τ_A becomes simply $\operatorname{var}(X) - \operatorname{var}(T_X)$, which is the variance of the error component of *X*. Therefore, Assumption 2 of generalized Levine equating (population invariance of the conditional variance of *X* given τ_A) becomes identical to Assumption 2 of Levine equating (population invariance of the error component of *X*). A similar result holds for *Y* and T_Y.

Poststratification Equipercentile Equating

This method, also known as *frequency estimation equipercentile equating*, is another classical equating method that applies to a NEAT design. (See Angoff, 1971, pp. 581–582; Braun & Holland, 1982, pp. 21–23; Kolen & Brennan, 2004, pp. 136–139). This method does not constrain the equating relationship to be a linear function. Assumptions 1 and 2 of the Tucker method are replaced by the assumption that the conditional distributions of *X* and of *Y*, given *A*, are population-invariant.

As before, the subscripts ***P*** and ***Q*** indicate the populations taking test forms *X* and *Y*; the subscript ***S*** indicates the synthetic population. *F*, *G*, and *H* represent cumulative distribution functions of scores *X*, *Y*, and *A*. In many cases, the distributions indicated by *F* and *G* will be conditional distributions, conditioning on the anchor score. For example, $F_{\boldsymbol{P}}(x|a)$ will represent the conditional distribution function of *X*, given $A = a$, in population ***P***.

The distribution function of anchor score *A* in population ***S*** is

$$H_{\boldsymbol{S}}(a) = wH_{\boldsymbol{P}}(a) + (1 - w)H_{\boldsymbol{Q}}(a). \tag{4}$$

Then the distribution of *X* in population ***S***, is

$$F_{\boldsymbol{S}}(x) = w\int F_{\boldsymbol{P}}(x|a)dH_{\boldsymbol{P}}(a) + (1-w)\int F_{\boldsymbol{Q}}(x|a)dH_{\boldsymbol{Q}}(a). \tag{5}$$

The assumption that the conditional distribution of *X* given $A = a$ is population-invariant makes it possible to substitute the conditional distributions in population ***P*** for the corresponding distributions in population ***Q***, leading to an estimate for the distribution of *X* in population ***S***,

$$\begin{aligned}\hat{F}_{\boldsymbol{S}}(x) &= w\int F_{\boldsymbol{P}}(x|a)dH_{\boldsymbol{P}}(a) + (1-w)\int F_{\boldsymbol{P}}(x|a)dH_{\boldsymbol{Q}}(a)\\ &= \int F_{\boldsymbol{P}}(x|a)dH_{\boldsymbol{S}}(a),\end{aligned} \tag{6}$$

where $H_{\boldsymbol{S}}(a)$ is given by Equation 4.

The assumption that the conditional distribution of *Y* given $A = a$ is population-invariant yields a similar estimate $\hat{G}_{\boldsymbol{S}}(y)$ for the distribution of *Y* in population ***S***. The poststratification equipercentile equating of *X* to *Y* is the equipercentile equating from $\hat{F}_{\boldsymbol{S}}(x)$ to $\hat{G}_{\boldsymbol{S}}(y)$.

Curvilinear Levine Observed Score Equating

Curvilinear Levine observed score equating (Chen & Holland, 2009) requires a transformation of the bivariate distributions of test scores and anchor scores—a mean-preserving linear transformation. The mean-preserving linear transformation with parameters *λ* and *ν* transforms (*X*, *A*) to a new pair of random variables (*X′*, *A′*),

$$\begin{aligned}X' &= \mu_X + \lambda(X - \mu_X)\\ A' &= \mu_A + \nu(A - \mu_A),\end{aligned} \tag{7}$$

where μ_X and μ_A are the means of X and A, respectively, and λ and v are positive real numbers.

The joint distribution of X' and A' has the same means as the joint distribution of X and A, as well as the same correlation coefficient, but the standard deviations of X and A are multiplied by the factors λ and v.

Chen and Holland (2009) noted that, in kernel equating (von Davier, Holland, & Thayer, 2004), if the distributions are continuized with a very large bandwidth, poststratification equipercentile equating (PSEE) becomes nearly identical to Tucker equating. They then showed that applying the mean-preserving linear transformation, with an appropriate choice of λ and v, to the joint distributions of (X, A) and of (Y, A) and then continuizing these joint distributions with a very large bandwidth will make PSEE nearly identical to Levine observed score equating. On the basis of that result, they defined curvilinear Levine observed score equating as the process of transforming the (X, A) and (Y, A) distributions with these values of λ and v and then using the transformed bivariate distributions to do poststratification equipercentile equating.

Poststratification Equipercentile Equating Based on True Anchor Scores

As the name suggests, this method (abbreviated PSEE_{TA}) is a variation of PSEE in which the anchor scores are replaced by their corresponding true scores, which is the same way that Levine observed score equating differs from Tucker equating. The conditional distributions of the tests X and Y, given the anchor true score, are assumed to be population-invariant. Using τ_a to represent a given true score on the anchor, one can estimate the cumulative distribution of X on population $\boldsymbol{S}$ as

$$\hat{F}_{\boldsymbol{S}}\left(x;\tau_A\right) = \int F_{\boldsymbol{P}}\left(x \mid \tau_a\right) dH_{\boldsymbol{S}}\left(\tau_a\right), \qquad (8)$$

where $F_{\boldsymbol{P}}\left(x \mid \tau_a\right)$ is the conditional distribution function of X given $T_A = \tau_a$ in population $\boldsymbol{P}$, and $H_{\boldsymbol{S}}\left(\tau_a\right)$ is the distribution function of T_A in population $\boldsymbol{S}$. The cumulative distribution of Y in population $\boldsymbol{S}$ is estimated similarly by $\hat{G}_{\boldsymbol{S}}\left(y;\tau_A\right)$, and the equating is the equipercentile equating from $\hat{F}_{\boldsymbol{S}}\left(x;\tau_A\right)$ to $\hat{G}_{\boldsymbol{S}}\left(y;\tau_A\right)$.

Results

In this section it will be shown that, under certain conditions, the linear form of PSEE$_{\text{TA}}$ is Levine observed score equating, and that the formula for curvilinear Levine observed score equating (Chen & Holland, 2009) can be used for PSEE$_{\text{TA}}$.

First, two terms need to be defined. They are the conditional mean of X given the true anchor score,

$$\mu_{\boldsymbol{P}}(X|\tau_a) = \int x dF_{\boldsymbol{P}}(x|\tau_a), \qquad (9)$$

and the conditional variance of X given the true anchor score:

$$\text{var}_{\boldsymbol{P}}(X|\tau_a) = \int [x - \mu_{\boldsymbol{P}}(X|\tau_a)]^2 dF_{\boldsymbol{P}}(x|\tau_a). \qquad (10)$$

The corresponding terms for Y are defined similarly.

Theorem

If in population $\boldsymbol{P}$, $\mu_{\boldsymbol{P}}\left(X \mid \tau_a\right)$ is a linear function of τ_a, and $\text{var}_{\boldsymbol{P}}\left(X \mid \tau_a\right)$ is constant on T_A, and if $\mu_{\boldsymbol{Q}}\left(Y \mid \tau_a\right)$ and $\text{var}_{\boldsymbol{Q}}\left(Y \mid \tau_a\right)$ have the same properties in population $\boldsymbol{Q}$, then the linear equating from $\hat{F}_{\boldsymbol{S}}\left(x;\tau_A\right)$ to $\hat{G}_{\boldsymbol{S}}\left(y;\tau_A\right)$ is generalized Levine observed score equating.

The proof is given in Appendix A.

Corollary

Additionally, if both T_X and T_Y are correlated perfectly with T_A, and the variance of the error component of A is population-invariant, then the linear equating from $\hat{F}_{\boldsymbol{S}}\left(x;\tau_A\right)$ to $\hat{G}_{\boldsymbol{S}}\left(y;\tau_A\right)$ is Levine observed score equating.

Theoretically, PSEE$_{\text{TA}}$ can be regarded as the curvilinear analogue to Levine observed score equating. However, this equating has no practical value unless one can estimate the distributions of (X, T_A) and (Y, T_A). Some data models can produce the joint distribution of (X, T_A) or (Y, T_A), but the fitting will sometimes fail the statistical tests for large size samples. One approach is to use the mean-preserving linear transformation of Equation 7 to transform (X, A) and (Y, A) into (X', A') and (Y', A'), choosing the values of λ_X, ν_X, λ_Y, and ν_Y that make the linear PSEE (i.e., Tucker equating) based on (X', A') and (Y', A') equal to the linear PSEE$_{\text{TA}}$ (i.e., Levine observed score equating) based on the original datasets (*X*, *A*) and (*Y*, *A*). Hence, the curvilinear form of PSEE$_{\mathrm{TA}}$ based on the original datasets (*X*, *A*) and (*Y*, *A*) can be approximated with the curvilinear form of PSEE based on (*X*′, *A*′) and (*Y*′, *A*′).

One way to choose the values of the *λ*s and *ν*s is to set a linear equating from *X*′ to *Y*′,

$$y \;=\; \mu_S(Y') + \frac{\sigma_S(Y')}{\sigma_S(X')}[x - \mu_S(X')], \tag{11}$$

where

$$\mu_S\left(X'\right) \;= \int \mu_{\boldsymbol{P}}\left(X' \mid a'\right) dH_S\left(a'\right), \tag{12}$$

$$\mu_S\left(Y'\right) \;= \int \mu_{\boldsymbol{Q}}\left(Y' \mid a'\right) dH_S\left(a'\right), \tag{13}$$

$$\sigma_S^2\left(X'\right) \;= \int \mathrm{var}_{\boldsymbol{P}}\left(X' \mid a'\right) dH_S\left(a'\right) + \int \mu_{\boldsymbol{P}}^2\left(X' \mid a'\right) dH_S\left(a'\right) - \mu_S^2\left(X'\right), \tag{14}$$

and

$$\sigma_S^2\left(Y'\right) \;= \int \mathrm{var}_{\boldsymbol{Q}}\left(Y' \mid a'\right) dH_S\left(a'\right) + \int \mu_{\boldsymbol{Q}}^2\left(Y' \mid a'\right) dH_S\left(a'\right) - \mu_S^2\left(Y'\right). \tag{15}$$

Then set the term on the left side of each of these equations (12, 13, 14, and 15) equal to the corresponding term in the Levine observed score equating from *X* to *Y* on ***S***. One needs to adjust the marginal distributions of both *X* and *Y* to solve Equations 12, 13, 14, and 15, although in general the values of the *λ*s are close to 1. The formulas for the *λ*s and *ν*s are given in Appendix B.

Discussion

The idea for PSEE$_{\mathrm{TA}}$ came from observing the similarity between two equatings computed from the same data: a direct linear equating of distributions estimated by item response theory (IRT) and Levine observed score equating based on the same IRT-fitted data. (For the results of this comparison, see Chen, 2010.) It was then apparent that PSEE$_{\mathrm{TA}}$ provides a more natural definition for curvilinear Levine observed score equating than the method previously given that name (Chen & Holland, 2009), which only provides a computational procedure for approximating PSEE$_{\mathrm{TA}}$.

Several issues related to the results in this paper suggest directions for further research.

First, it is problematic that the mean-preserving linear transformation is applied to the bivariate distribution of two observed score variables to simulate the bivariate distribution of an observed score variable and a true score variable. The process requires the computation of σ_{T_X}, σ_{T_A}, and so on, which in turn assumes a perfect correlation between the true scores T_X and T_A and some additional conditions on (X, A). These assumptions are often violated by nonlinearity between X and A. Hence, the ratio of σ_{T_X}/σ_X, σ_{T_A}/σ_A, and so forth can be off target by 1% to 5%, which sometimes can cause a difference in the equating results as large as 0.15 or even 0.2 standard deviations. This violation of the assumption of a perfect true-score correlation between test and anchor scores explains why the Levine method is often not as accurate as other linear methods. Even if by chance σ_{T_X}, σ_{T_A}, and so on are estimated perfectly, the true score distribution estimated by applying the mean-preserving linear transformation procedure on the observed score distribution can be quite different from its actual distribution. It can be shown that when an IRT model is used to estimate the joint distributions of test scores and anchor scores, PSEE_{TA} is quite different from PSEE with the mean-preserving linear transformation technique, particularly at the end-score points. Therefore, to make a reliable PSEE_{TA} procedure, it is necessary to develop a model that can produce the joint distribution for two variables—observed scores on the test and true scores on the anchor—and that can fit the data well for the majority of data sets. Some progress has been made in this direction.

The second issue is more critical. In equating through an anchor, without IRT, the choice of an equating method involves a judgment as to the extent to which the assumptions of each method are likely to be satisfied by the data. Based on the data, one can choose the Tucker method, the Levine method, the chained linear method, their curvilinear versions, or other methods. On the other hand, IRT equating, with any IRT model, is actually poststratification equipercentile equating, stratifying on true anchor scores (Chen, 2010), which we now know is essentially the curvilinear form of the Levine method. Therefore, IRT equating, like Levine observed-score equating, will tend to make too large an adjustment for the ability difference between populations $\boldsymbol{P}$ and $\boldsymbol{Q}$. Should one adopt the IRT viewpoint so that only Levine-type equating methods can be used? Or should one try to choose an equating method on a case-by-case basis, and if so, by what criteria? Certainly, many more theoretical and technical questions are waiting to be answered.

Conclusion

In this paper, poststratification equipercentile equating, stratifying on true anchor scores (PSEE_{TA}) is defined, and its linear form is shown to be the Levine observed score equating method. Hence, this method, rather than the method defined in Chen and Holland (2009), is the method that should be called the curvilinear Levine observed score equating method.

Based on the results in Chen, Livingston, and Holland (2010), most, if not all, of the commonly used equating methods for NEAT designs can be approximated as either (a) observed anchor score-based poststratification equating, which includes PSEE, Tucker, and Braun-Holland (Braun & Holland, 1982); or (b) true anchor score-based poststratification equating, which includes both Levine methods, hybrid Levine equipercentile equating, chained true score equipercentile equatings, curvilinear Levine observed score equating, and IRT-based equating; or (c) partially true anchor score-based poststratification equating, which includes chained equipercentile equating, chained linear equating, and modified poststratification equating (Wang & Brennan, 2007). Theoretically, the definition of modified poststratification equating is the same as that of PSEE_{TA}. Computationally, both methods use the mean-preserving linear transformation to modify the existing distributions, but modified poststratification equating estimates only the marginal distributions of the true anchor scores, while PSEE_{TA} estimates the joint distribution of observed test scores and true anchor scores.

Two important problems remain to be solved. The first problem is to develop a model for estimating a joint distribution of observed test scores and true anchor scores—one that can fit the data better than currently used latent variable models. The second problem is to develop a criterion (or a set of criteria) for determining which method—PSEE, PSEE_{TA}, or some other equating method—is most appropriate for a given equating task.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55–69). New York, NY: Academic Press.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9–49). New York, NY: Academic Press.

Chen, H. (2010, May). *A comparison between linear IRT observed score equating and Levine observed score equating.* Paper presented at the annual meeting of the National Council of Measurement in Education, Denver, CO.

Chen, H., & Holland, P. W. (2008, March). *Construction of chained true score equipercentile equatings under the KE framework and their relationship to Levine true score equating* (Research Report No. RR-09-24). Princeton, NJ: Educational Testing Service.

Chen, H., & Holland, P. W. (2009, April). *Nonlinear Levine observed score equating*. *Or is it?* Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Chen, H., Livingston, S. A., & Holland, P. W. (2010). Generalized equating functions for NEAT designs. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling and linking* (pp. 185–200). New York, NY: Springer-Verlag.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Rao, C. R. (1973). *Linear statistical inference and its applications*. New York, NY: Wiley.

von Davier, A. A., Fournier-Zajac, S., & Holland, P. W. (2006, April). *An equipercentile version of the Levine linear observed-score equating function using the methods of kernel equating*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer-Verlag.

Wang, T., & Brennan, R. L. (2007, April). *A modified frequency estimation equating method for the common-item non-equivalent groups design*. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Notes

[1] See Angoff, 1982; Braun & Holland, 1982.

Appendix A

Proof of the Main Theorem

One needs to show that Assumptions A1 and A2 are satisfied for both X and Y defined from PSEE_{TA} with the conditions in Theorem 1.

Let $\mu_{\boldsymbol{P}}(X|\tau_a)$ defined in Equation 9 be a linear function of τ_a, that is,

$$\mu_{\boldsymbol{P}}(X|\tau_a) = \alpha * \tau_a + \beta, \tag{A1}$$

where both α and β will be determined later. Sum up $\mu_{\boldsymbol{P}}(X|\tau_a)$ on T_A with population $\boldsymbol{P}$, using both Equations 9 and A1, one has

$$\mu_{\boldsymbol{P}}(X) = \alpha * \mu_{\boldsymbol{P}}(T_A) + \beta. \tag{A2}$$

Then taking the mean of $\mu_{\boldsymbol{P}}(X|\tau_a)$ on T_A with population $\boldsymbol{P}$, using both Equations 9 and A1 again, there is

$$\mu_{\boldsymbol{P}}(X{*}T_A) = \alpha * \mu_{\boldsymbol{P}}(T_A{}^2) + \beta * \mu_{\boldsymbol{P}}(T_A). \tag{A3}$$

Both Equations A2 and A3 are the same equations for solving the regression of X on T_A in population $\boldsymbol{P}$. Hence, if the regression of X on T_A in population $\boldsymbol{Q}$ has the same α and β, it can be said that the regression is population-invariant.

By assuming that the conditional distribution of X given $T_A = \tau_a$ is population-invariant, that is, $F_{\boldsymbol{Q}}(x|\tau_a) = F_{\boldsymbol{P}}(x|\tau_a)$, one can define $\mu_{\boldsymbol{Q}}(X|\tau_a)$ in the form of Equation 9 as

$$\mu_{\boldsymbol{Q}}(X|\tau_a) = \int x d\boldsymbol{F}_{\boldsymbol{Q}}(x|\tau_a) = \int x d\boldsymbol{F}_{\boldsymbol{P}}(x|\tau_a) = \alpha * \tau_a + \beta. \tag{A4}$$

Hence, the regression of X on T_A is population-invariant. This proves that Assumption A1 is valid for X.

From Equations A2 and A3, one gets:

$$\alpha = \frac{\operatorname{cov}(X, T_A)}{\sigma^2_{\boldsymbol{P}}(T_A)}, \tag{A5}$$

and

$$\beta = \mu_{\boldsymbol{P}}(X) - \frac{Cov(X, T_A)}{\sigma^2_{\boldsymbol{P}}(T_A)} \mu_{\boldsymbol{P}}(T_A). \tag{A6}$$

With the assumption in the theorem that $\mathrm{var}_{\boldsymbol{P}}(X|\tau_a)$ = c is a constant on T_A, the integration of $\mathrm{var}_{\boldsymbol{P}}(X|\tau_a)$ over T_A in population ***P*** gives

$$\begin{aligned} c &= \mu_{\boldsymbol{P}}[\mathrm{var}_{\boldsymbol{P}}(X \mid T_A)] \\ &= \mathrm{var}_{\boldsymbol{P}}(X) - \mathrm{var}_{\boldsymbol{P}}[\mu_{\boldsymbol{P}}(X \mid T_A)] \text{ (Law of total variance)} \\ &= \mathrm{var}_{\boldsymbol{P}}(X) - \mathrm{var}_{\boldsymbol{P}}(\alpha * T_A + \beta) \\ &= \mathrm{var}_{\boldsymbol{P}}(X) - \alpha^2 \mathrm{var}_{\boldsymbol{P}}(T_A) \\ &= \mathrm{var}_{\boldsymbol{P}}(X) - \frac{\mathrm{cov}^2_{\boldsymbol{P}}(X, T_A)}{\mathrm{var}^2_{\boldsymbol{P}}(T_A)} \mathrm{var}_{\boldsymbol{P}}(T_A) \\ &= \mathrm{var}_{\boldsymbol{P}}(X)[1 - \rho^2_{\boldsymbol{P}}(X, T_A)]. \end{aligned} \tag{A7}$$

Here the law of total variance is used (see Rao, 1973, p. 97, Equation (2b.3.6); see also Equations A1 and A5 in this paper.)

Similar to Equation 10, $\mathrm{var}_{\boldsymbol{Q}}(X|\tau_a)$ can be defined as

$$\mathrm{var}_{\boldsymbol{Q}}(X|\tau_a) = \int [x - \mu_{\boldsymbol{Q}}(X|\tau_a)]^2 dF_{\boldsymbol{Q}}(x|\tau_a). \tag{A8}$$

With the assumption that $F_{\boldsymbol{Q}}(x|\tau_a) = F_{\boldsymbol{P}}(x|\tau_a)$ and the result that $\mu_{\boldsymbol{Q}}(X|\tau_a) = \mu_{\boldsymbol{P}}(X|\tau_a)$, one can see that $\mathrm{var}_{\boldsymbol{Q}}(X|\tau_a)$ = c is a constant on T_A as well. Integrating $\mathrm{var}_{\boldsymbol{Q}}\ (X|\tau_a)$ over T_A in population ***Q*** gives

$$c = \mathrm{var}_{\boldsymbol{Q}}(X)[1 - \rho^2_{\boldsymbol{Q}}(X, T_A)]. \tag{A9}$$

This proves that Assumption A2 is valid for *X* also. Similarly, both assumptions can be shown to be valid for *Y*.

Appendix B

Formulas for Parameter Values in Equations 12–15

Given (X, A) and (Y, A) in a NEAT design, let

$$X' = \mu_X + \lambda_1(X - \mu_X), \text{ and } A' = \mu_{A_{\boldsymbol{P}}} + v_1(A - \mu_{A_{\boldsymbol{P}}}) \text{ on } \boldsymbol{P}, \tag{B1}$$

and

$$Y' = \mu_Y + \lambda_2(Y - \mu_Y), \text{ and } A' = \mu_{A_{\boldsymbol{Q}}} + v_2(a - \mu_{A_{\boldsymbol{Q}}}) \text{ on } \boldsymbol{Q}, \tag{B2}$$

respectively. Where the subscripts to A indicate which population is used, the proper values to make so-called distributions based on true anchor scores are:

$$\lambda_1 = \sqrt{\frac{1 - w\rho^2(Y,A) - (1-w)\dfrac{\rho_X^2}{\rho_{A_P}^2}\left[1 + \dfrac{\sigma_{A_Q}^2}{\sigma_{A_P}^2}\left(\dfrac{\rho^2(Y,A)\rho_{A_Q}^2}{\rho_Y^2} - 1\right)\right]}{1 - w\rho^2(Y,A) - (1-w)\rho^2(X,A)}}, \tag{B3}$$

$$\nu_1 = \frac{\rho(X,A)\rho_{A_P}}{\rho_X}\lambda_1, \tag{B4}$$

$$\lambda_2 = \sqrt{\frac{1 - (1-w)\rho^2(X,A) - w\dfrac{\rho_Y^2}{\rho_{A_Q}^2}\left[1 + \dfrac{\sigma_{A_P}^2}{\sigma_{A_Q}^2}\left(\dfrac{\rho^2(X,A)\rho_{A_P}^2}{\rho_X^2} - 1\right)\right]}{1 - (1-w)\rho^2(X,A) - w\rho^2(Y,A)}}, \tag{B5}$$

and

$$\nu_2 = \frac{\rho(Y,A)\rho_{A_Q}}{\rho_Y}\lambda_2. \tag{B6}$$

where ρ_X is $\sigma_{T_X} / \sigma_{X'}$ and so forth.

In general, λ_1 and λ_2 are close to 1 (Chen & Holland, 2009).