Lecture 13: Test Equating

Equating is the statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences in test form difficulty, not for differences in content.

The goal in test equating is to determine a correspondence between scores on a test X and a test Y that matches comparable scores. A test equating function is used to take a score x on test X and determine a score y^* on test Y that would make the examinee obtaining score x comparable to examinees administered test Y.

Classical Methods of Equating (not based on IRT)

These methods determine a correspondence between scores on X and Y by matching the score distributions for populations administered X and Y. In most applications of these methods, it is assumed that both tests are administered to randomly equivalent populations of examinees, so that the only differences between the resulting test score distributions are due to differences between the tests, not the people who took the tests.

- Equipercentile equating – matches scores that have the same percentile rank.

$$e_Y(x) = G^{-1}(F(x))$$

  where F is the cumulative distribution function (c.d.f.) for X and G is the c.d.f. for Y.

- Linear equating – defines a linear transformation of scores on X to scores on Y such that both test score distributions have the same mean and variance.

$$e_Y(x) = \sigma_Y \frac{x - \mu_X}{\sigma_X} + \mu_Y.$$

Equating with Item Response Theory

Lord (1980) defines a condition of *equity* which says that two scores are equated provided it is a matter of indifference to examinees at every ability level which of the two tests is taken. In other words,

$$G^*(e_Y(x)|\theta) = G(y|\theta) \text{ at all } \theta.$$

Lord's equity property implies that examinees with a given true score would have identical observed score means, standard deviations, and distributional shapes of converted scores on Form X and scores on Form Y.

Some implications of this condition:

- Tests measuring different traits cannot be equated.
- Tests that have unequal reliabilities cannot be equated.
- Only tests that are perfectly reliable or strictly parallel can be equated in a way that satisfies exactly the equity condition.

Because this last condition can never be satisfied, the real question to ask is not whether an equating function results in equity, but how close to equity the equating function brings us. Sometimes Lord's equity condition is referred to as a full equity condition because it requires that the conditional distributions be exactly the same at all θs.

Some less-restrictive equity criteria include:

- First-order or Weak Equity: The expected scores for examinees on tests X and Y (once equated) are the same at all ability levels [$E_X(e_Y(x)|\theta) = E_Y(Y|\theta)$ at all θ].

- Second-order Equity: The means and variances of test scores examinees on tests X and Y (once equated) are the same at all ability levels [in addition to first-order equity; also $Var_X(e_Y(x)|\theta) = Var_Y(Y|\theta)$ at all θ]

Lord considers two other important properties of an equating function:

- Symmetry: The equating function should be the same whether equating X to Y or Y to X. This property requires that the function used to transform a score on Form X to the Form Y scale be the inverse of the function used to transform a score on Form Y to the Form X scale.

- Invariance across groups: The equating function should be the same regardless of the subpopulation used to determine it.

The following is a list of steps for implementing equating (the order might vary in practice) – pp. 7-8.

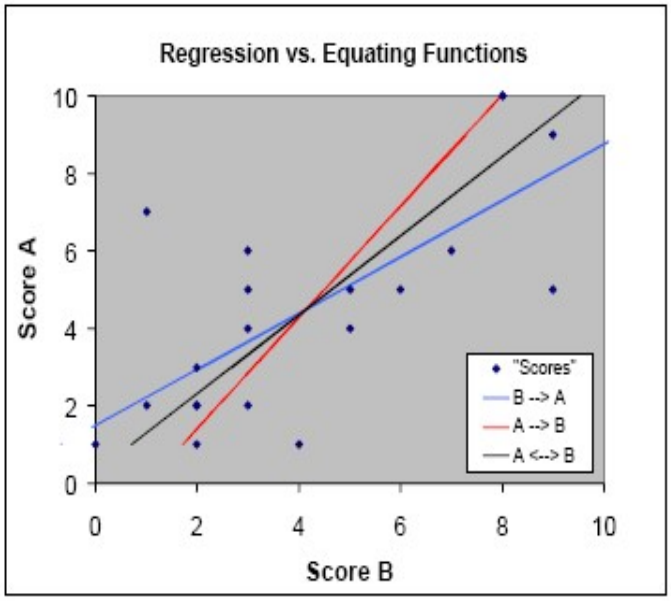
1. Decide on the purpose for equating.
2. Construct alternate forms. Alternate test forms are constructed in accordance with the same content and statistical specification.
3. Choose a design for data collection. Equating requires that data be collected for providing information on how the test forms differ statistically.
4. Implement the data collection design. The test is administered and the data are collected as specified by the design.
5. Choose one or more operational definitions of equating. Equating requires that a choice be made about what types of relationships between forms are to be estimated. For

example, this choice might involve deciding on whether to implement linear or nonlinear equating methods.

6. Choose one or more statistical estimation methods. Various procedures exist for estimating a particular equating relationship.
7. Evaluate the results of equating. After equating is conducted, the results need to be evaluated.

Prediction vs. Equating

- Score equating is not prediction (e.g, regression). Both approaches transform scores on one test into the scale of the scores on another test.
- Different purposes: In prediction, the goal is to predict an expected Y-score for an examinee from some other information about that examinee. In prediction, there is an inherent asymmetry. Equating functions do not predict scores on one test from scores on another. Instead, scores that have been equated can be used interchangeably. Symmetry is essential.



Scaling in IRT

As we have already seen, when the ability parameters are known, the item parameter estimates are invariant. This means that regardless of which examinees are used, the item parameter estimates should stay the same. Likewise, if the item parameters are known, ability estimates are on the same scale even when examinees are administered different items. When neither is known, however, abilities and item parameters estimated independently are only invariant up to a linear transformation of the θ-scale. The process of finding an appropriate transformation such that both sets of parameters are on the same scale is referred to as scaling. Suppose A and B denote two different tests whose item parameters where estimated separately. Then the item parameters will be related by the following relationship:

$$b_A = mb_B + d \qquad (1)$$

$$a_A = \frac{a_B}{m} \qquad (2)$$

$$c_A = c_B \qquad (3)$$

and the ability parameters by the following relationship:

$$\theta_A = m\theta_B + d \qquad (4)$$

But in order to determine the slope m and intercept d parameters of this linear transformation, we need to have a link that somehow connects the two calibrations.

Linking Designs

- Single Group design – the same group of examinees are administered each of the two different tests
- Equivalent Groups design – the two tests are given to groups that can be assumed to be equivalent
- Anchor Test design – the two tests are given to different groups of examinees, but there is a subset of the items that stay the same across groups

- Common Persons design – some but not all of the examinees are administered both test forms

Determining the Scaling Constants *m* and *d*:

- Regression method – for items or examinees that are the same across administrations, equations (1) to (3) could be estimated using ordinary least-squares regression. The problem with this type of transformation is that it does not satisfy a symmetry condition.

- Mean/Sigma method – Let S_A and S_B be the standard deviations and $\bar{b}_A$ and $\bar{b}_B$ the means of the b_As, b_Bs for the common items across tests A and B, respectively. Then m is estimated as $\frac{S_B}{S_A}$ and d as $\bar{b}_B - m\bar{b}_A$.
- Mean/Mean method – The mean/mean method is the same as the mean/sigma method, but uses the mean of the a parameters to determine m. Then, the mean of the *b*-parameter estimates of the common items is used in place of the parameters to estimate

the *d* constant. The values of *m* and *d* then can be substituted into (1) to (4) to obtain the rescaled parameter estimates.

- Characteristic Curve method – for the full set of anchor items, determine test characteristic curves based on the item parameter estimates from the two separate calibrations.

1) Haebara (1980) approach – expresses the difference between the ICCs as the sum of the squared difference between the ICCs for each item for examinees of a particular ability.

$$Hdiff(\theta_j) = \sum_{i:V} \left[p_{ij}(\theta_{Aj}; \hat{a}_{Ai}, \hat{b}_{Ai}, \hat{c}_{Ai}) - p_{ij}(\theta_{Aj}; \frac{\hat{a}_{Bi}}{m}, m\hat{b}_{Bi} + d, \hat{c}_{Bi}) \right]^2$$

$Hdiff$ is cumulated over examinees. The estimation process proceeds by finding *m* and *d* that minimizes the following criterion:

$$Hcrit = \sum_j Hdiff(\theta_j)$$

2) Stocking & Lord (1983) approach – uses the square of differences of the sums,

$$SLdiff(\theta_j) = \left[\sum_{i:V} p_{ij}(\theta_{Aj}; \hat{a}_{Ai}, \hat{b}_{Ai}, \hat{c}_{Ai}) - \sum_{i:V} p_{ij}(\theta_{Aj}; \frac{\hat{a}_{Bi}}{m}, m\hat{b}_{Bi} + d, \hat{c}_{Bi}) \right]^2$$

The summation is taken over items for each set of parameter estimates before squaring. $SLdiff(\theta_j)$ then is cumulated over examinees. The estimation proceeds by finding the combination of *m* and *d* that minimizes the following criterion:

$$SLcrit = \sum_j SLdiff(\theta_j)$$

*** Equating Recipes Open-source Code and Monograph & Equating/Linking Programs are available at http://www.education.uiowa.edu/centers/casma/computer-programs

*** https://cran.r-project.org/web/packages/SNSequate/SNSequate.pdf

https://cran.r-project.org/web/packages/equateIRT/equateIRT.pdf

https://cran.r-project.org/web/packages/irtoys/irtoys.pdf

https://cran.r-project.org/web/packages/equate/equate.pdf