Workshop

Test Equating

Jorge Tendeiro, Andrea Stoevenbelt

24 May 2017

university of groningen

Overview

1. What is equating?
2. Desirable properties of equating
3. Equating designs
4. Error in equating analyses
5. Equating under the random groups design (classic methods)
6. Equating under the NEAT design (classic methods)
7. Equating using IRT
8. Standard errors of equating
9. Choosing among equating methods
10. Sample size requirements
11. Empirical example

Main idea

- Tests are widespread assessment tools.
- Different test forms are often administered on multiple occasions.
- Test forms are usually built under strict test specifications.
- Test specifications try to ensure that test forms are as similar as possible.
- But, are they?

"This year's exam was waaay easier than last year, no?"

"The resit was sooo much harder than the regular exam! Why?!?"

Main idea

How to make sure that test forms are similar?
$\longrightarrow$ Perform test equating.

Definition

Equating is a 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 difficulty among forms that are built to be similar in difficulty and content.

(Kolen & Brennan, 2014, p. 2)

Main idea

Example

Student A took the regular Stats exam in 2015-16: Grade $= 5.7 \rightarrow 6$.
Student B took the regular Stats exam in 2016-17: Grade $= 5.2 \rightarrow 5$.

Possible explanations:

- Student A is higher achieving than Student B.
- The 2015-16 exam was easier, in spite of the exams being 'parallel' (unfair!).

Equating accounts for differences in test forms' difficulty by converting (say) the grade of the 2016-17 exam onto the grade scale of the 2015-16 exam.

- After equating, *differences in grades are no longer attributed to differences in exam difficulty*.

Caveats

- Equating adjusts for differences in difficulty, not for differences in content.

  Test forms are expected to be as similar as possible in content and statistical specifications. (Otherwise, consider linking.)

- Depending on the testing setting and on the test equating model, some assumptions need to be met in order for equating to proceed.

- In some circumstances, not equating may be the best solution.
  (E.g., if samples are extremely small.)

Desirable properties of equating

Symmetry *The equating function is invertible.*
Example: If $f_{New \to Old}(6.2) = 6.6$, then $f_{Old \to New}(6.6) = 6.2$.
(In particular, regression is ruled out!)

Same specifications *Test forms are similar in content and statistical specifications.*

(Weak) Equity *It does not matter which test form a student takes.*
Examinees are expected to earn the same (equated) score on the NEW form as they would on the OLD form.

Desirable properties of equating

Symmetry *The equating function is invertible.*
Example: If $f_{New \to Old}(6.2) = 6.6$, then $f_{Old \to New}(6.6) = 6.2$.
(In particular, regression is ruled out!)

Same specifications *Test forms are similar in content and statistical specifications.*

(Weak) Equity *It does not matter which test form a student takes.*
Examinees are expected to earn the same (equated) score on the NEW form as they would on the OLD form.

Some assumptions are method-specific:

Mean equating The distribution of the converted scores of Form_{New} has the same mean as the distribution of scores of Form_{Old}.

Linear equating The distribution of the converted scores of Form_{New} has the same mean and SD as the distribution of scores of Form_{Old}.

Equipercentile equating The distribution of the converted scores of Form_{New} is equal to the distribution of scores of Form_{Old}.

Equating designs

Three common designs:

Random groups design Forms are administered simultaneously (one form per examinee, randomly assigned).

Equating designs

Three common designs:

Random groups design Forms are administered simultaneously (one form per examinee, randomly assigned).

Single group design with counterbalancing Forms are administered simultaneously to all examinees (in alternating order).

Equating designs

Three common designs:

Random groups design Forms are administered simultaneously (one form per examinee, randomly assigned).

Single group design with counterbalancing Forms are administered simultaneously to all examinees (in alternating order).

Common-item nonequivalent groups design (NEAT) Test forms have common items, which are used to control for ability differences between the groups of examinees.

The common items:

- Are a 'mini version' of the total test form.
- Behave similarly in both forms (e.g., similarly placed).
- Should be ipsis verbis the same.
- Are either internal or external.
- Adjust for group (i.e., population) differences.

The NEAT design is by far the most popular design.

Error in equating analyses

Two types of error exist:

- Random equating error

  Due to sampling from population of examinees to estimate parameters.
  Quantifiable (standard error of equating).
  N larger $\Rightarrow$ random equating error smaller.

Error in equating analyses

Two types of error exist:

- Random equating error

  Due to sampling from population of examinees to estimate parameters.
  Quantifiable (standard error of equating).
  N larger $\Rightarrow$ random equating error smaller.

- Systematic equating error

  Due to the equating method used (and failure to meet its assumptions).
  For example, under the NEAT design, systematic error may be expected when:

  - The common items are not representative of the whole test.
  - The common items function differently for each group of examinees.

  Hard to quantify.
  Larger samples don't necessarily reduce the problem.

In cases where large equating error is to be expected, not equating may be preferable to equating.

Equating under the random groups design (classic methods)

Assume the two basic equating properties (symmetry, same specifications).
Three methods:

- Mean equating.
- Linear equating.
- Equipercentile equating.

Notation:

Form X	New form
Form Y	Old form
$eq_Y(x)$	Equating function that converts scores on Form X to the scale of Form Y.

Mean equating

Assumption

Form X is considered to differ in difficulty from Form Y by a constant amount along the score scale.

$$x - \mu(X) = y - \mu(Y),$$

so

$$eq_Y(x) = y = \underbrace{1}_{A} x + \underbrace{[\mu(Y) - \mu(X)]}_{B}.$$

$$= Ax + B, \text{a straight line}$$

Linear equating

Assumption

Scale scores differ in both location and scale.

$$\frac{x-\mu(X)}{\sigma(X)} = \frac{y-\mu(Y)}{\sigma(Y)},$$

so

$$eq_Y(x) = y = \underbrace{\frac{\sigma(Y)}{\sigma(X)}}_{A} x + \underbrace{\left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right]}_{B}.$$

$$= Ax + B, \text{a straight line}$$

Mean and linear equating properties

- For both the mean and the linear equating methods,

$$\mathbf{E}[eq_Y(X)] = \mu(Y).$$

The mean of Form X scores equated to the Form Y scale is equal to the mean of the Form Y scores.

- For mean equating,

$$\sigma[eq_Y(X)] = \sigma(X).$$

The SD of Form X scores equated to the Form Y scale is equal to the SD of the Form X scores.

- For linear equating,

$$\sigma[eq_Y(X)] = \sigma(Y).$$

The SD of Form X scores equated to the Form Y scale is equal to the SD of the Form Y scores.

- Both the mean and linear equating methods are symmetric, i.e.,

$$eq_Y(x) = y \Rightarrow eq_X(y) = x.$$

Equipercentile equating

A curvilinear function (instead of a straight line) is estimated. So, more general than mean or linear equating.

Example

Equipercentile (but not mean or linear equating) can be used if Form X is more difficult than Form Y at high and low scores, but less difficult at the middle scores.

Goal

The distribution of scores on Form X converted to the Form Y scale is equal to the distribution of scores on Form Y, based on percentile ranks:

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

where

- $F(x) = P(X \leq x)$
- $G(y) = P(Y \leq y)$

Equipercentile equating

Example

- Suppose that, on Form X, 30% of students scored 6.0 or less (i.e., $P(X \leq 6) = .30$).
- Suppose that, on Form Y, 30% of students scored 6.5 or less (i.e., $P(Y \leq 6.5) = .30$).

Then

$$eq_Y(x = 6.0) = 6.5.$$

Equipercentile equating

Example

- Suppose that, on Form X, 30% of students scored 6.0 or less (i.e., $P(X \leq 6) = .30$).
- Suppose that, on Form Y, 30% of students scored 6.5 or less (i.e., $P(Y \leq 6.5) = .30$).

Then

$$eq_Y(x = 6.0) = 6.5.$$

Notes

Equipercentile equating:

- Does meet the symmetry assumption.
- Assumes that test scores are continuous random variables. For integer test scoring some adaptations are needed.
- For (much) more details go to Kolen and Brennan (2014).

Equipercentile equating – Smoothing

- Sample percentiles are typically associated to large SEs.
- Consequently, empirical distributions are imprecise (very 'bumpy').
- True even when sample size seems large (e.g., a few 1000s!).

Smoothing allows approximating empirical distributions (and equipercentile relationships) based on the main distribution trend, removing sample-based irregularities.

Risk

Smoothed empirical distributions may be poor approximations of the true (population) distribution (i.e., systematic equating error).

Goal

- Achieve more precise and stable equating relations (not so sample-dependant).
- Strike a balance between accuracy and parsimony.
- Thus, the reduction of random error is expected to offset the possible introduction of systematic error.

Equipercentile equating – Smoothing

Two types of smoothing:

Presmoothing Smooth the score distributions first and then equate.

Postsmoothing Equate first, and then smooth the equipercentile equated scores.

Popular smoothing methods include:

- Polynomial log-linear presmoothing.
- Cubic spline postsmoothing.
- Kernel smoothing (not so efficient as the previous two).
- Strong true score method (requires specification of distributional form of true scores). Lord's *beta4* method.

Equating under the NEAT design (classic methods)

Several methods, linear (L) and nonlinear (NL):

- Tucker method (L).
- Levine observed score method (L).
- Levine true score method (L).
- Chained linear equating (L).
- Equipercentile methods (NL): (Modified) frequency estimation method, chained equipercentile equating.

Notation:

Form X	New form, taken by Population 1
Form Y	Old form, taken by Population 2
V	Common-item set (anchor items)
$eq_Y(x)$	Equating function that converts scores on Form X to the scale of Form Y.

Equating under the NEAT design (classic methods)

General form of the linear equating function:

$$eq_Y(x) = y = \underbrace{\frac{\sigma_s(Y)}{\sigma_s(X)}}_{A} x + \underbrace{\left[\mu_s(Y) - \frac{\sigma_s(Y)}{\sigma_s(X)}\mu_s(X)\right]}_{B}.$$

$$= Ax + B, \text{a straight line}$$

The subscript '$_s$' denotes a synthetic population, which is a weighted population derived from Population 1 and Population 2:

$$w_1 + w_2 = 1, \text{ with } w_1, w_2 \geq 0.$$

Equating under the NEAT design (classic methods)

General form of the linear equating function:

$$eq_Y(x) = y = \underbrace{\frac{\sigma_s(Y)}{\sigma_s(X)}}_{A} x + \underbrace{\left[\mu_s(Y) - \frac{\sigma_s(Y)}{\sigma_s(X)}\mu_s(X)\right]}_{B}.$$

$$= Ax + B, \text{a straight line}$$

The subscript '$_s$' denotes a synthetic population, which is a weighted population derived from Population 1 and Population 2:

$$w_1 + w_2 = 1, \text{ with } w_1, w_2 \geq 0.$$

- Mathematically, the difference between the Tucker, the Levine observed score, the Levine true score, and the chained linear methods is the computational formulas for A and B above (not shown here).
- Statistically, different equating methods are based on different model assumptions.

Equating using IRT

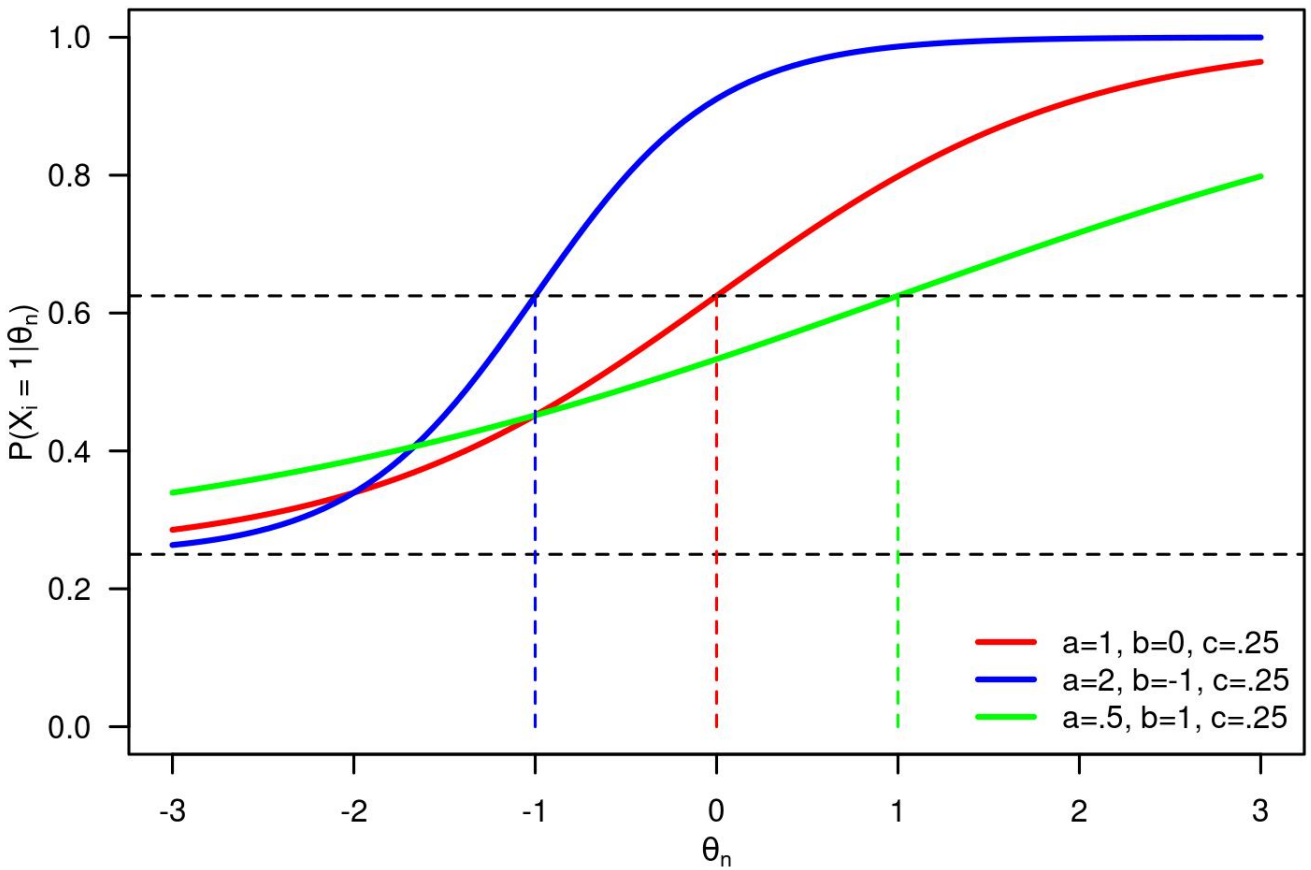
Three steps are typically followed:

1. Fit IRT model to data.
2. Linearly transform IRT parameter estimates to a base scale.
3. Convert scores on Form X to the scale of Form Y.

Equating using IRT – Step 1 (Fit IRT model to data)

- For dichotomous scores use the common 1PLM, 2PLM, or 3PLM.
- Assumptions: Unidimensionality, monotonicity, local independence.

$$P(X_i = 1|\theta_n) = c_i + \frac{1 - c_i}{1 + \exp\left[-a_i(\theta_n - b_i)\right]}$$



Equating using IRT – Step 2 (Transform IRT parameters)

Fact: The 3PLM model is invariant under linear transformation of θ.
Consequence: IRT parameter estimates from different test forms are on different IRT scales. This is relevant under the NEAT design.

Equating using IRT – Step 2 (Transform IRT parameters)

Fact: The 3PLM model is invariant under linear transformation of θ.
Consequence: IRT parameter estimates from different test forms are on different IRT scales. This is relevant under the NEAT design.

Example: Suppose $A = .5$, $B = 2$.

Form X					Form Y			
a_i	b_i	θ_n	c_i	$P_i(\theta_n)$	a_i/A	$A \times b_i + B$	$A \times \theta_n + B$	c_i
1	0	1	.25	.80	2	2	2.5	.25
2	1	−1	.20	.21	4	2.5	1.5	.20
1.6	−1	0	.18	.86	3.2	1.5	2	.18

Equating using IRT – Step 2 (Transform IRT parameters)

Fact: The 3PLM model is invariant under linear transformation of θ.
Consequence: IRT parameter estimates from different test forms are on different IRT scales. This is relevant under the NEAT design.

Example: Suppose $A = .5$, $B = 2$.

Form X					Form Y			
a_i	b_i	θ_n	c_i	$P_i(\theta_n)$	a_i/A	$A \times b_i + B$	$A \times \theta_n + B$	c_i
1	0	1	.25	.80	2	2	2.5	.25
2	1	−1	.20	.21	4	2.5	1.5	.20
1.6	−1	0	.18	.86	3.2	1.5	2	.18

- Step 2 takes care of removing this IRT scale difference between item parameter estimates from different forms.
- Find suitable constants A, B that do the job.
  (The methods available include: Mean/sigma, mean/mean, Haebara, Stocking-Lord transformations.)
- This is called calibration.

Equating using IRT – Step 3 (Convert scores X $\rightarrow$ Y.)

- Surprisingly, θ estimates ($\widehat{\theta}$) from the 3PLM are not used directly to perform equating. Reasons:
  - The number-correct (NC) score is not a sufficient statistic for θ. That is, different response patters with the same NC score may imply different θ values. This is hard to convey to test takers.
  - $\widehat{\theta}$ is difficult to get (unlike NC scores).
  - The precision of $\widehat{\theta}$ is low on the extremes (and high in the middle).
- NC scores are commonly used instead.

Methods available to convert NC scores include:

- IRT true score equating – Based on IRT's NC true score:

$$\tau_X(\theta_n) = \sum_i P(X_i = 1|\theta_n; a_i, b_i, c_i)$$

- IRT observed score equating – Equipercentile equating based on IRT-derived distributions of observed NC scores on each form.

Standard errors of equating

- The focus is on random error (due to sampling), not on systematic error (due to the equating method used).
- N increases $\Rightarrow$ random error decreases.
- Bootstrap is the method of choice to estimate random error, although analytic formulas also exist.
- As to be expected,
  Standard error of equating (SEE) = SD of the equating parameter of interest over many replications of the equating procedure (on random samples of equal size).
- SEEs are conditional on scores on Form X.

Choosing among equating methods

- There is no uniformly best equating method. Many factors influence the choice of the 'best' equating model.
- Below are some rough guidelines.

Identity equating

Very (too) small samples, similar test form difficulties, assumptions of other methods bluntly violated.

Mean, linear equating

Small sample size, similar test form difficulties, precision close to mean values required.

Nonlinear equating (equipercentile, 3PLM IRT)

Large(ish) sample sizes, test forms can differ in difficulty, common items need to be representative of entire test, relationship not linear, precision along all score scale.

Sample size requirements

Simple rules of thumb (random groups and NEAT designs):

- ~ 400 per form for linear equating and 1PLM IRT equating
- $\sim 1,500$ per form for equipercentile equating and 3PLM IRT equating

For really small samples (say, < 100), specific methods have been developed:

- Equipercentile equating with smoothing (Livingston, 1993).
- Circle-arc equating (Livingston & Kim, 2009).
- Synthetic link function equating (Kim, von Davier, & Haberman, 2008).
- Equating using collateral information (Livingston & Lewis, 2009).
- Nominal weights equating (Babcock, Albano, & Raymond, 2012).

Empirical example

Analysis done by Andrea Stoevenbelt (manually and using 'equate' in R).

Data Biopsychology exam 2013, 2015.

Design Nonequivalent groups common items (NEAT).
80 items per exam, 34 common items.

Equating methods used: Linear only (Tucker, Levine observed score, Levine true score, chained linear equating).

Empirical example

Analysis done by Andrea Stoevenbelt (manually and using 'equate' in R).

Data Biopsychology exam 2013, 2015.

Design Nonequivalent groups common items (NEAT).
80 items per exam, 34 common items.

Equating methods used: Linear only (Tucker, Levine observed score, Levine true score, chained linear equating).

Descriptive statistics

Group	N	I	Score	$\widehat{\mu}$	$\widehat{\sigma}$
2013	349	46	Total score	63.07	10.23
		34	Common-items score	26.25	5.18
2015	384	46	Total score	63.71	8.37
		34	Common-items score	26.72	4.52

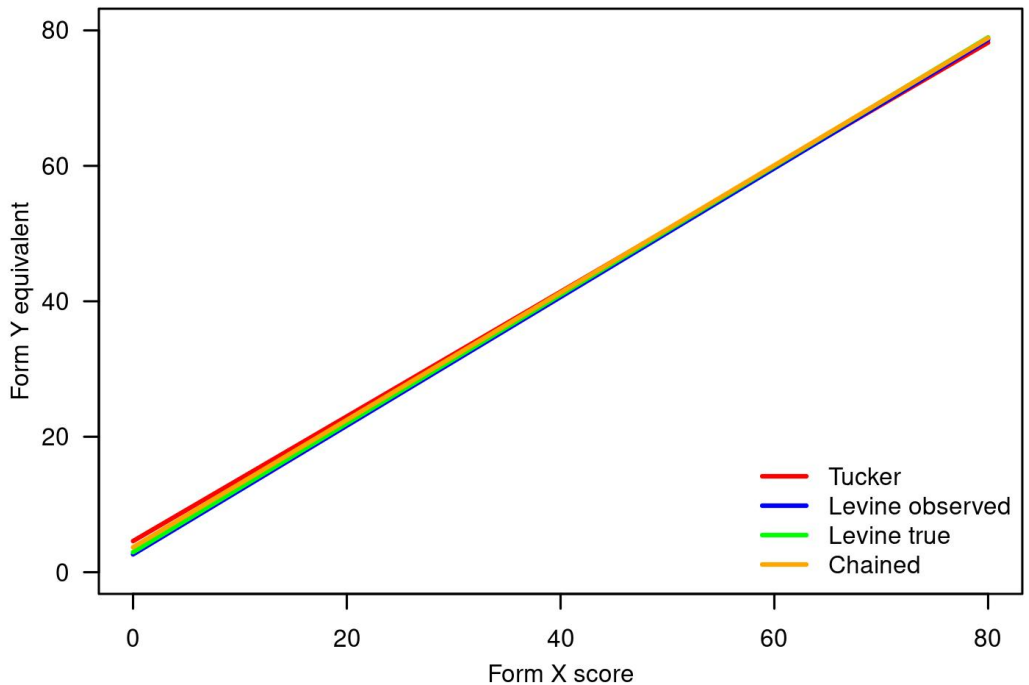
Note: As the descriptives suggest, hardly any equating seems needed in this case.

Empirical example

Results linear equating

Equating method	Intercept	Slope
Tucker	4.60	0.92
Levine observed score	2.64	0.95
Levine true score	2.96	0.95
Chained linear equating	3.68	0.94

The linear equating functions are virtually indistinguishable.



Empirical example

Some conclusions:

- Very small differences between scores on 2015 exam and equivalent 2013 scores.
- Equated scores for $X \leq 30$ were unreliable due to extrapolation.

For more information...

Albano, A. D. (2016). equate: An R package for observed-score linking and equating. *Journal of Statistical Software*, *74*, 1-36.

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Greenwood, Westport.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices (3rd Ed.)*. Springer Verlag: New York, NY.