This is the published version of a paper published in .

Citation for the original published paper (version of record):

equateIRT Package in R
https://doi.org/10.1080/15366367.2018.1492866

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-151901
equateIRT Package in R

Marie Wiberg

To cite this article: Marie Wiberg (2018) equateIRT Package in R, Measurement: Interdisciplinary Research and Perspectives, 16:3, 195-202, DOI: 10.1080/15366367.2018.1492866

To link to this article: https://doi.org/10.1080/15366367.2018.1492866

© 2018 The Author(s). Published by Taylor & Francis.

Published online: 10 Sep 2018.

Article views: 48

View Crossmark data
SOFTWARE REVIEW

equateIRT Package in R

Marie Wiberg

Department of Statistics, USBE, Umeå University

ABSTRACT
Equating test scores between different achievement test versions is important to assure comparability between test takers' scores. As many items are modelled with item response theory (IRT), it makes sense to also equate the test scores with IRT equating methods. The equateIRT package in R provides a set of functions which implements IRT equating methods including newer extensions. This paper summarizes some of the advances in equating with IRT, reviews the equateIRT package, and demonstrates, through two illustrative examples, some of the key features of the package.

KEYWORDS
IRT observed-score equating; IRT true-score equating; equating coefficients; bisector method; IRT

Introduction
When different test versions are used to measure the same ability, it is important to have methods to ensure that the test takers can be compared regardless of which test version they have taken. Equating refers to a family of statistical models and methods that are used to make test scores comparable among different test versions so that the scores can be used interchangeably (González & Wiberg, 2017). There are many different equating methods available depending on the data collection design and what assumptions are made. If you model items with item response theory (IRT; Lord, 1980) a common tool for test constructors when creating and analyzing tests, it makes sense to use IRT equating methods when equating test scores.

To perform IRT equating, one can use the equateIRT package in R (Battauz, 2015), which provides a set of commands that implement traditional IRT equating methods as well as newer extensions. The aim of this article is to summarize some of the advances in equating with IRT, review the equateIRT package, and demonstrate, through two illustrative examples, the capability of the package. In the next section, some theory of equating with IRT will be summarized, then a brief description of the equateIRT features is given. Next, two illustrative examples are given and the paper ends with a conclusion that provides some limitations and suggestions for the future.

Equating with IRT
Assume we have a test form $g$, then the probability that a test taker with ability $\theta_g$ answers an item correctly can be modelled with the three-parameter logistic (3PL) IRT model

$$
\pi_{g_i}(\theta_g, a_{g_i}, b_{g_i}, c_{g_i}) = c_{g_i} + \left(1 - c_{g_i}\right) \frac{\exp(Da_{g_i}(\theta_g - b_{g_i}))}{1 + \exp(Da_{g_i}(\theta_g - b_{g_i}))},
$$

where $a_{g_i}$ is the item discrimination, $b_{g_i}$ is the item difficulty, and $c_{g_i}$ is the pseudo item guessing in test form $g$. $D$ is a constant which is commonly set to 1.7. If we set $c_{g_i} = 0$, we get the 2PL IRT model and also if $a_{g_i} = 1$, we get the 1PL IRT model (Lord, 1980). In order to equate two different test
forms, we need to set the parameter estimates on the same scale, which is done with the help of equating coefficients.

Equating coefficients

Let \( n_g \) be the number of items in test form \( g \) and \( n_g - 1 \) be number of items in common with another test form \( g - 1 \). By using the equating coefficients \( A_{g-1,g} \) and \( B_{g-1,g} \), the parameters estimated from test form \( g - 1 \) can be transformed to the scale of test form \( g \) as follows:

\[
a_g = \frac{a_{g-1}}{A_{g-1,g}}, \quad b_g = A_{g-1,g}b_{g-1} + B_{g-1,g}, \quad \text{and} \quad \theta_g = A_{g-1,g}\theta_{g-1} + B_{g-1,g}.
\]

The equating coefficients can be estimated either from methods based on response functions or moments of item parameters (Kolen & Brennan, 2014. Chaps 6.3.2–6.3.3). Methods implemented in equateIRT include the response function methods of Haebara and Stocking-Lord and the moment-based methods; mean-mean, mean-geometric mean, and the mean-sigma. \( B_{g-1,g} \) is estimated in the same way for all moment-based methods and is given by

\[
B_{g-1,g} = \frac{1}{n_{g-1,g}} \sum_{i=1}^{n_{g-1,g}} b_{gi} - A_{g-1,g} \frac{1}{n_{g-1,g}} \sum_{i=1}^{n_{g-1,g}} b_{g-1,gi}, \tag{1}
\]

where \( A_{g-1,g} \) is method dependent. For the mean-sigma method, which will be used in the later illustrative examples, it is defined by

\[
A_{g-1,g} = \sqrt{\sum_{i=1}^{n_{g-1,g}} b_{gi}^2 - \frac{1}{n_{g-1,g}} \left( \sum_{i=1}^{n_{g-1,g}} b_{gi} \right)^2} \left( \sum_{i=1}^{n_{g-1,g}} b_{g-1,gi}^2 - \frac{1}{n_{g-1,g}} \left( \sum_{i=1}^{n_{g-1,g}} b_{g-1,gi} \right)^2 \right)^{-\frac{1}{2}}. \tag{2}
\]

The asymptotic covariance matrix for the vector of estimates of the equating coefficients can be derived with the delta method for the equating coefficients and is summarized in Battauz (2015).

The equating coefficient described so far is used when we have a situation where we can use a direct equating. For more details, refer to Battauz (2015).

If two test forms are linked through different paths, it is possible to average the equating relationship to obtain a single transformation which is expected to be more accurate (Kolen & Brennan, 2014, p. 280). In the equateIRT package, the bisector method for equating (Battauz, 2013) yields a weighted average of linear transformations \( \theta_k = \sum_{p \in P_0} w_p \theta_k^p \), where \( \theta_k^p = A_p \theta_0 + B_p \) represent
the transformation of $\theta_0$ to $\theta_k$, $w_p = n_p(1 + A^2_p)^{-1/2}/\left(\sum_{b \in P_{na}} n_b(1 + A^2_b)^{-1/2}\right)$, and $n_p$ are optional path weights. With the bisector method, the equating coefficients are defined as

$$A_{0k}^* = \sum_{p \in P_{na}} A_p w_p \quad \text{and} \quad B_{0k}^* = \sum_{p \in P_{na}} B_p w_p.$$ 

Again, it is easy to obtain the asymptotic covariance matrix of the estimates and this feature is implemented in the equateIRT package.

In large-scale assessments, it is common to use several test versions which are linked together through a series of anchor items, i.e. an indirect equating. In order to link several test versions, one typically uses a linkage plan and the feature to see which test forms have common items is implemented in the equateIRT package.

### Equating

There are two different equating methods, which are implemented in the equateIRT package: IRT true-score equating (TSE) and IRT observed-score equating (OSE). IRT TSE (Lord, 1980) uses the mean of the conditional score distributions and uses the assumed true score a test taker has. An observed score can be defined as the true score plus an error term. The idea is to equate a true score $\tau_X$ associated with a given ability $\theta$ on test form $X$ with the true score $\tau_Y$ on another test form $Y$. The equating transformation can thus be defined as

$$\varphi(\tau_X, \theta) = T^{Y}(T^{X^{-1}}(\tau_X)),$$

where

$$\tau_X = E(X|\theta) = \sum_{j \in X} \pi_{ij}(\theta, \varpi_j) = T^{TX}(\theta),$$

$$\tau_Y = E(Y|\theta) = \sum_{j \in Y} \pi_{ij}(\theta, \varpi_j) = T^{TY}(\theta),$$

and where $\varpi_j$ is a vector of item characteristics (e.g. item difficulty or item discrimination) (González & Wiberg, 2017).

IRT OSE (Lord, 1980) uses the marginal score distributions and IRT models are used to define the involved conditional score probabilities. First, one assumes distributions for the test takers abilities and these distributions are integrated or summed across ability levels to obtain marginal observed-score distributions for test forms $X$ and $Y$. Then, equipercentile equating is applied to these distributions as follows.

$$\varphi(x) = F_Y^{-1}(F_X(x)),$$

where $F_X$ and $F_Y$ are the cumulative distribution functions which can be obtained by using either Lord and Wingersky (1984) algorithm or through other approaches discussed in González, Wiberg, and von Davier (2016).

### The R package equateIRT

The R package equateIRT has implemented functions to estimate the equating coefficients and their corresponding standard errors (SEs) using the previous mentioned methods. The package allows you to get an overview of which test forms have common items and thus give you the linkage plan. If you have two test forms with common items, it can perform direct equating and if you have pairs of common items on multiple test forms, it allows you to use indirect or chain equating. The equateIRT package has both IRT OSE and IRT TSE implemented for dichotomous items. If you have several possible equating paths, you can get a single transformation by averaging the equating coefficients.
with the bisector method. The R package supports the Rasch, 1PL, 2PL, and 3PL IRT models but it does not estimate the item parameters nor their covariance matrices. Instead, equateIRT allows you to import estimates of the item parameters and the covariance matrices from flexMIRT (Cai, 2013), IRTPRO (Cai, Thissen, & du Toit, 2011), and the R packages ltm (Rizopoulos, 2006) and mirt (Chalmers, 2012). To import item parameter estimates and the covariance matrix, one can use the functions import.ltm(), import.mirt(), import.flexmirt(), and import.irtpro(). The imported data are then used in the equating to obtain analytical SEs for direct, chain, and average equating coefficients. In the next two subsections, two illustrative examples will demonstrate some of the key features of the equateIRT package.

**Illustrative equating example**

To illustrate how to use equateIRT to perform IRT OSE and IRT TSE, we will use two test forms from a binary scored college admissions test where each form contains 40 common anchor items in the first columns and 80 unique items in the following columns. The used data are freely available and can be downloaded using the provided links. We will start by fitting a 2PL IRT model and obtain the item parameter estimates and their SEs with the R package mirt. The obtained item parameter estimates and SEs are stored in the objects mADMx.2PL and mADMy.2PL which are then read into equateIRT with the import.mirt() function as follows.

```r
> library(mirt)
> mADMx.2PL <- mirt(ADMneatX, 1, itemtype = "2PL", SE = TRUE)
> mADMy.2PL <- mirt(ADMneatY, 1, itemtype = "2PL", SE = TRUE)
> library(equateIRT)
> estX.2PL <- import.mirt(mADMx.2PL, display = FALSE)
> estY.2PL <- import.mirt(mADMy.2PL, display = FALSE)
```

As equateIRT requires the anchor items to have the same names in the test forms to be equated, we renamed the 40 common items as c1, . . ., c40 and give the unique items the names X1, . . ., X80 and Y1, . . ., Y80, respectively. Then, we create lists of matrices with the item parameter estimates and the covariances and name the two test forms “Test1” and “Test2”.

```r
> a.x <- as.matrix(estX.2PL$coef)
> a.y <- as.matrix(estY.2PL$coef)
> row.names(aux.x) <- c(paste("c", 1:40, sep = ""), paste("X", 1:80, sep = ""))
> row.names(aux.y) <- c(paste("c", 1:40, sep = ""), paste("Y", 1:80, sep = ""))
> estXY.2PL <- list(a.x, a.y)
> estXYvar <- list(estX.2PL$var, estY.2PL$var)
> tests <- paste("Test", 1:2, sep = "")
```

Next, item parameter linking should be done to place the item parameter estimates on a common scale using the functions modIRT(coef, var = NULL, names = NULL...) and direc(). The argument coef reads in the list of matrices containing the item parameter estimates, and the argument var is used to read in the covariance matrix of item parameter estimates if it is available. The option names are used if we want to give specific names to the test forms. The obtained modIRT object contains a list of the item parameter estimates and their covariance matrix (if it was entered) for each of the test forms one wants to link. The modIRT object is then read into direc() to perform the linking step.

In the direc(mods, which, method = "mean-mean", ...) function, mods is an object of class modIRT which contains item parameter coefficients and their covariance matrix of the forms to be equated. The statement “which” tells the program which test forms to equate. Finally, method states which equating method is used and the alternatives are “mean-mean”, “mean-sigma”, “mean-gmean”, “Haebara”, or “Stocking-Lord”. If one provides the covariance matrix, the column StdErr in the output gives you the SE of the equating coefficients A and B. If one has not provided the
covariance matrix of the item parameter estimates, the output shows NA. We illustrate these two functions using the mean-sigma method with equating coefficients defined in Equations (1) and (2).

```r
> m2plXY <- modIRT(coef = estXY.2PL, var = estXYVar, names = tests, display = FALSE)
> t12 <- direc(mods = m2plXY, which = c(1,2), method = "mean-sigma")
> summary(t12)
```

```
Link: Test1.Test2
Method: mean-sigma
Equating coefficients:
  Estimate StdErr
  A 0.8832732 0.091612
  B 0.0024206 0.038019
```

Finally, the score(obj, method = , se = , w = 0.5) function is used to obtain the equated scores and their SEs using either IRT OSE or IRT TSE. For each of the methods, we display the equated scores for the last five values (scores = 76:80) when using the weight w = 1 in the synthetic population. First, we give the codes for IRT OSE, where the columns in the output from left to right show the scores on Test 2 and the equated scores and their SEs (StdErr).

```r
> score(t12, method = "OSE", se = TRUE, scores = 76:80, w=1)
  Test1 as.Test2 StdErr
  77 76 76.06792 1.633756
  78 77 77.13020 1.684269
  79 78 78.17833 1.714616
  80 79 79.22108 1.734992
  81 80 80.26698 1.754759
```

Second, we display the code for IRT TSE, where the columns are similar as for IRT OSE except for the first column which gives the estimated theta values which correspond to the equated scores.

```r
> score(t12, method = "TSE", se = TRUE, scores = 76:80, w=1)
  theta Test2 Test1.as.Test2 StdErr
  1 0.7651143 76 76.05460 1.605256
  2 0.8168637 77 77.14264 1.685284
  3 0.8693834 78 78.23046 1.763214
  4 0.9227257 79 79.31797 1.838789
  5 0.9769456 80 80.40505 1.911753
```

**Illustrative example linkage plans and indirect equating**

To illustrate linkage plans and chained equating or indirect equating, we use the five data sets in data2pl which comes with the equateIRT package. First, we estimate a 2PL IRT model for the five data sets with the R package mirt. The estimated item parameters and covariances are then read into equateIRT with the import.mirt() function. Below we give the lines for the first data set but the other four are estimated similarly with mirt and imported similarly into equateIRT with the names est2, est3, est4, and est5.

```r
> data("data2pl", package = "equateIRT")
> library(mirt)
> m1 <- mirt(data2pl[[1]],1,itemtype = "2PL", SE=TRUE)
> library(equateIRT)
> est1 <- import.mirt(m1, display = FALSE)
```

Next, we create a list of coefficients and covariance matrices, name the test forms “test1-test5” and create an object of class modIRT. Similarly to the first numerical illustration we use as input in modIRT() the item parameter estimates and the covariance matrices.
The five data sets have different items in common and to get an overview of the linkage plan, we can use the linkp(coef) function which calculates the number of common items between a list of test forms.

```r
> linkp(coef = estC)

[1,] 20 10 0 0 10
[2,] 10 20 10 0 0
[3,] 0 10 20 10 0
[4,] 0 0 10 20 10
[5,] 10 0 0 10 20
```

From the output, we can see that test form 1 has 10 common items with test form 2 and test form 5. In order to estimate the direct equating coefficients and SEs between test forms 1 and 5, we would use the direc() function with the mean-sigma method.

```r
> dir15 <- direc(mods = m2pl, which = c(1,5), method = "mean-sigma")
> summary(dir15)

Link: test1.test5
Method: mean-sigma
Equating coefficients:
  Estimate StdErr
  A 1.0043 0.033101
  B -0.4931 0.027203
```

Another feature in equateIRT is the possibility to calculate all direct equating coefficients and SEs using IRT methods between all pairs of test forms with common items by using the function alldirec(mods = , method = , . . .). The function takes as input mods which is an object of class modIRT containing the item parameter coefficients and their covariance matrix of the forms to be equated. Similar to the direc() function you can decide which equating method to use and here we illustrate the function with the mean-sigma method.

```r
> direclist1 <- alldirec(mods = m2pl, method = "mean-sigma")
> direclist1

Direct equating coefficients
Method: mean-sigma
Links:
test1.test2
test1.test5
test2.test1
test2.test3
test3.test2
test3.test4
test4.test3
test4.test5
test5.test1
test5.test4
```

Another possibility is to use the function chainec(r = NULL, direclist, f1 = NULL, f2 = NULL, pths = NULL) to estimate all chain (indirect) equating coefficients and SEs using IRT methods. The function allows you to specify the length of the chain that is the number of forms used for equating (r). It takes as input direclist which is an object returned from the alldirec function which contains direct equating coefficients between pairs of test forms as seen above. You can also specify the
starting test form (f1) and the ending test form (f2) or the specific equating path you prefer to use (pths). First, we illustrate the function by

estimating all chain equating coefficient of length \( r = 3 \) from test form 2 to test form 5.

```r
> cec25 <- chainec(r = 3, direclist = direclist1, f1 = "test2", f2 = "test5")
> summary(cec25)
Path: test2.test1.test5
Method: mean-sigma
Equating coefficients:
    Estimate StdErr
A  0.84341 0.046263
B -0.37425 0.027976
```

If we want to estimate the chain equating coefficient for the specific path \( \{2,3,4,5\} \), we can write code as follows.

```r
> pth <- paste("test", c(2,3,4,5), sep = "")
> chainec2345 <- chainec(direclist = direclist1, pths = pth)
> summary(chainec2345)
Path: test2.test3.test4.test5
Method: mean-sigma
Equating coefficients:
    Estimate StdErr
A  0.93421 0.057406
B -0.47365 0.039411
```

If one has different equating paths, the package provides the option to calculate average equating coefficients using the bisector method and SEs given a set of direct and chain equating coefficients using the function `bisectorec(ecall = , mods = NULL, weighted = TRUE, unweighted = TRUE)`. The function takes as input `ecall`, which is a list of objects of the classes returned from either the function `direc()` or `chainec()`. The option `weighted` is logical and if true weighted bisector coefficients are estimated, likewise the logical option `unweighted` if TRUE computes unweighted bisector coefficients.

```r
> ecall <- c(cec25, chainec2345)
> av25 <- bisectorec(ecall = ecall, weighted = TRUE, unweighted = TRUE)
> summary(av25)
Link: test2.test5
Method: mean-sigma
Equating coefficients:
Path Estimate StdErr
A test2.test1.test5 0.84341 0.046263
A test2.test3.test4.test5 0.93421 0.057406
A bisector 0.88779 0.037406
A weighted bisector 0.87968 0.036854
B test2.test1.test5 -0.37425 0.027976
B test2.test3.test4.test5 -0.47365 0.039411
B bisector -0.42283 0.027424
B weighted bisector -0.41395 0.026510
```

The output gives the equating coefficients and their SEs for different paths, if they are averaged with the bisector method and if a weighted bisector method is used.

**Conclusions**

The equateIRT package contains a number of suitable features for conducting IRT equating, between both two test forms and when we have a more complicated linkage plan as is common in large-scale assessments. The package is also used as import in the R package kequate (Andersson, Bränberg, & Wiberg, 2013) where it is used when performing IRT observed-score kernel equating. The package is very useful with only minor limitations. It does not allow the user to use mixed item types. A possible
way around this is to use a general model and constrain some of the parameters. Another limitation is that it only allows you to equate two forms, although with possible different paths. The two form equating limitation is resolved using the R package equateMultiple (Battauz, 2017) which allows the calculation of equating coefficients between multiple test forms. A third limitation might be that only analytical SEs are implemented. It would be useful if bootstrap SEs were offered as an option within the package. These are minor limitations, which can be solved through use of other packages.

The overall conclusion is that equateIRT is easy to use, flexible, and is a very useful package if researchers or test constructors want to perform equating with IRT.

Funding
This work was supported by the Swedish Research Council grant 2014-578.

References
Battauz, M. (2017). equateMultiple: Equating of multiple forms. R package version 0.0.0. Downloaded from cran.r-project.org on June 18.