

Package ‘transDIF’

October 21, 2026

Title Score Comparability for Translated and Adapted Exams

Version 0.2.0

Description An adaptation-comparability workflow for small, lower-scoring and unbalanced language groups where standard differential item functioning (DIF) tools (Magis, Beland, Tuerlinckx and De Boeck, 2010, <[doi:10.3758/BRM.42.3.847](https://doi.org/10.3758/BRM.42.3.847)>) break down. Calibrates the Rasch model in each language group, links the groups robustly through the densest cluster of items rather than assuming DIF cancels out (on anchor selection see Kopf, Zeileis and Strobl, 2015, <[doi:10.1177/0013164414529792](https://doi.org/10.1177/0013164414529792)>), detects small-sample DIF with an empirical-Bayes spike-and-slab model and local false discovery rates (Efron, 2004, <[doi:10.1198/016214504000000089](https://doi.org/10.1198/016214504000000089)>), quantifies whether item-level DIF accumulates into different pass rates, explains DIF by item features to give translators actionable guidance, and drafts a comparability report.

License MIT + file LICENSE

URL <https://github.com/edidatasolutions/transDIF>,
<https://edidatasolutions.github.io/transDIF/>

BugReports <https://github.com/edidatasolutions/transDIF/issues>

Encoding UTF-8

Depends R (>= 4.1)

Imports stats

Suggests knitr, markdown

VignetteBuilder knitr

Config/roxygen2/version 8.1.0

NeedsCompilation no

Author Daniel Edi [aut, cre, cph] (ORCID:
<<https://orcid.org/0000-0001-5475-819X>>)

Maintainer Daniel Edi <danieledi2026@gmail.com>

Repository CRAN

Date/Publication 2026-10-21 03:10:02 UTC

Contents

td_calibrate	2
td_dif	3
td_features	4
td_impact	5
td_mh	6
td_report	6
td_sensitivity	7
td_simulate	9
Index	11

td_calibrate	<i>Calibrate each language group separately</i>
--------------	---

Description

Fits the Rasch model by marginal maximum likelihood within each group, each with its own mean-zero ability scale. The between-language difference ‘d = b_focal - b_ref’ therefore equals item DIF plus a common shift ‘c = -(focal mean ability on the reference scale)’. Separating the two is the linking problem [td_dif()] solves.

Usage

```
td_calibrate(responses, group, ref = "ref", focal = "focal")
```

Arguments

- responses 0/1 matrix (persons x items), column names = item ids.
- group Group label per person.
- ref, focal Labels of the reference (source-language) and focal (translated) groups.

Value

An ‘td_calibration’: data frame ‘items’ (‘item’, ‘b_ref’, ‘se_ref’, ‘b_focal’, ‘se_focal’, ‘d’, ‘se_d’) plus group sizes, ability SDs and ‘loc_var’ (variance of the two scales’ locations). ‘se_ref’/‘se_focal’ are absolute SEs; ‘se_d’ uses relative SEs, excluding the location uncertainty that is common to all items (it belongs to the linking shift).

Examples

```
sim <- td_simulate(n_ref = 400, n_focal = 120, n_items = 20, seed = 5)
cal <- td_calibrate(sim$responses, sim$group)
head(cal$items)
```

td_dif

*Robust linking, anchor selection and small-sample DIF***Description**

The between-language difference of item ‘i’ is modeled as $d_i \sim N(c + \delta_i, \text{se}_i^2)$, where ‘c’ is the common shift (the ability difference) that linking must recover and ‘ δ_i ’ is the item’s DIF.

Usage

```
td_dif(
  calibration,
  fdr = 0.1,
  tau0 = 0.05,
  anchor_max = 0.2,
  link = c("mode", "mixture"),
  ambiguity = 0.8,
  c_fixed = NULL,
  c_se_fixed = NA_real_
)
```

Arguments

calibration	A ‘td_calibration’.
fdr	Target Bayesian false discovery rate for flagging.
tau0	SD of DIF among "DIF-free" items: the scale of DIF considered negligible (logits).
anchor_max	Items with posterior DIF probability below this are reported as anchors.
link	"mode" (default) or "mixture".
ambiguity	Competing-mode density ratio above which the linking is reported as ambiguous.
c_fixed	Optional linking shift chosen by the analyst (for example, purified or all-item linking); the DIF model is then fitted given it. Used by [td_sensitivity()].
c_se_fixed	Standard error to attach to ‘c_fixed’.

Details

****Linking.**** By default ‘c’ is the precision-weighted *mode* of the ‘ d_i ’: the center of the densest cluster of items, found by maximizing $\sum_i \phi((d_i - c) / s_i) / s_i$ with $s_i^2 = \text{se}_i^2 + \text{tau0}^2$. It assumes that DIF-free items form the largest cluster, not that DIF cancels out as mean linking does. In known-truth benchmarks across balanced, directional and heavy DIF, it had the lowest overall RMSE of the estimators compared (including mean linking, iterative purification, median, Tukey biweight, least trimmed squares and the joint mixture below). Its SE is a sandwich estimate plus the scales’ location variance. If a second cluster of items is nearly as dense, the linking is flagged as ambiguous and the competing shift is reported.

****DIF**** Given ‘c’, DIF effects follow a spike-and-slab mixture: DIF-free items (share ‘ $\pi_0 > 0.5$ ’) have ‘ $\delta_i \sim N(0, \tau_0^2)$ ’, and DIF items have ‘ $\delta_i \sim N(m_1, \tau_1^2)$ ’, where ‘ m_1 ’ allows directional DIF. Each item gets a posterior DIF probability, a shrunken DIF estimate and a local false discovery rate. ‘link = “mixture”’ instead estimates ‘c’ jointly in the mixture (the approach of version 0.1.0; kept for comparison).

Value

A ‘td_dif’ object: ‘\$items’ (‘item’, ‘d’, ‘se_d’, ‘p_dif’, ‘lfd_r’, ‘dif_mean’, ‘dif_sd’ (posterior mean/SD of DIF), ‘flag’, ‘anchor’), ‘\$link’ (‘c’, ‘c_se’, ‘pi0’, ‘m1’, ‘tau0’, ‘tau1’, ‘alt_c’, ‘alt_ratio’), baseline linking constants ‘c_mean’ (all items as anchors) and ‘c_purified’, and ‘weakly_identified’ (TRUE, with a warning, when a second item cluster is nearly as dense as the chosen one).

Examples

```
sim <- td_simulate(n_ref = 400, n_focal = 120, n_items = 20, seed = 5)
dif <- td_dif(td_calibrate(sim$responses, sim$group))
dif
# true linking shift is -focal_mean = 0.5
c(estimate = dif$link[["c"]], mean_linking = dif$c_mean)
```

 td_features | *Which item features predict DIF?* |

Description

Random-effects meta-regression of each item’s DIF estimate (‘d - c’) on item features (e.g. idioms, cultural referents, measurement units, vocabulary load), weighting by ‘ $1 / (se_d^2 + \tau^2)$ ’. The between-item variance ‘ τ^2 ’ not explained by the features is estimated by the method of moments. Coefficients are logits of DIF per unit of the feature. That is guidance a translation team can act on.

Usage

```
td_features(dif, features)
```

Arguments

dif	An ‘td_dif’.
features	Data frame with ‘item’ and numeric feature columns.

Value

Data frame of coefficients (‘term’, ‘estimate’, ‘se’, ‘z’, ‘p_value’) with attribute ‘tau’ (residual DIF SD). Features with no variation across items are dropped with a warning.

Examples

```
sim <- td_simulate(n_ref = 400, n_focal = 120, n_items = 20, seed = 5)
dif <- td_dif(td_calibrate(sim$responses, sim$group))
td_features(dif, sim$features)
```

td_impact

*Does item-level DIF add up to different pass rates?***Description**

Differential test functioning for the translated form. With difficulties on the reference scale ('b_ref') and DIF effects 'delta', the translated form has difficulties 'b_ref + delta'. For the translated-language population (ability $N(-c, \sigma_{\text{focal}}^2)$ on the reference scale) the function computes the pass rate at raw cut 'cut' on the translated form versus on a DIF-free form, using exact score distributions. It also computes the expected raw-score shift for an examinee exactly at the cut. Intervals propagate both the linking uncertainty (the focal group's mean ability is '-c') and the posterior uncertainty of each item's DIF.

Usage

```
td_impact(dif, cut, n_draws = 200, level = 0.9, seed = NULL)
```

Arguments

dif	An 'td_dif'.
cut	Raw-score passing standard.
n_draws	Posterior draws.
level	Interval level.
seed	Optional seed.

Value

A data frame with estimate and interval for 'pass_rate_fair', 'pass_rate_translated', 'pass_rate_change' and 'score_shift_at_cut'.

Examples

```
sim <- td_simulate(n_ref = 400, n_focal = 120, n_items = 20, seed = 5)
dif <- td_dif(td_calibrate(sim$responses, sim$group))
td_impact(dif, cut = 12, n_draws = 50, seed = 1)
```

td_mh	<i>Mantel-Haenszel DIF with purification (conventional baseline)</i>
-------	--

Description

Mantel-Haenszel DIF with purification (conventional baseline)

Usage

```
td_mh(responses, group, ref = "ref", focal = "focal", fdr = 0.1, purify = TRUE)
```

Arguments

responses	0/1 matrix.
group	Group label per person.
ref, focal	Group labels.
fdr	Benjamini-Hochberg level for flagging.
purify	Re-run once matching on the total over non-flagged items.

Value

Data frame: 'item', 'alpha_mh', 'delta_mh' (ETS delta scale), 'p_value', 'flag'.

Examples

```
sim <- td_simulate(n_ref = 400, n_focal = 120, n_items = 20, seed = 5)
mh <- td_mh(sim$responses, sim$group)
table(flagged = mh$flag, true_dif = sim$truth$dif_item)
```

td_report	<i>Draft a comparability report</i>
-----------	-------------------------------------

Description

Assembles a plain-language Markdown report of the linking, DIF, aggregate impact and feature analyses, organized around the kinds of evidence the Standards for Educational and Psychological Testing (2014; fairness chapter) and the ITC Guidelines for Translating and Adapting Tests (2nd ed., 2017) ask for. It is a draft for a psychometrician to review, not a finished validity argument.

Usage

```
td_report(
  dif,
  impact = NULL,
  features = NULL,
  languages = c("source", "target"),
  file = NULL,
  sensitivity = NULL
)
```

Arguments

<code>dif</code>	An ‘td_dif’.
<code>impact</code>	Optional output of [td_impact()].
<code>features</code>	Optional output of [td_features()].
<code>languages</code>	Names of the source and target languages.
<code>file</code>	Optional path to write the Markdown to.
<code>sensitivity</code>	Optional output of [td_sensitivity()]; adds a section on how the conclusions depend on the linking assumption.

Value

The report as a character string (invisibly if written to file).

Examples

```
sim <- td_simulate(n_ref = 400, n_focal = 120, n_items = 20, seed = 5)
dif <- td_dif(td_calibrate(sim$responses, sim$group))
rep <- td_report(dif, td_impact(dif, cut = 12, n_draws = 50, seed = 1),
  td_features(dif, sim$features), c("English", "French"),
  sensitivity = td_sensitivity(dif, B = 50, seed = 1))
cat(rep)
# To save: td_report(dif, file = file.path(tempdir(), "comparability.md"))
```

td_sensitivity	<i>Sensitivity of the comparability conclusions to the linking assumption</i>
----------------	---

Description

Linking two language groups requires an assumption about which items are free of DIF, and when DIF is pervasive on a short test the data may not settle it. This function makes that dependence visible:

Linking assumptions The ability difference under the weighted mode (DIF-free items form the densest cluster), iterative purification, and all items as anchors (DIF cancels out), with the DIF model refitted under each and, if ‘cut’ is given, the pass-rate impact under each.

Stability of the mode A parametric bootstrap redraws each item's between-language difference from its sampling distribution and re-estimates the mode. When DIF-free items form a clear cluster the bootstrap modes stay close to the estimate; when they do not, the mode jumps between clusters, which is exactly the failure a single estimate hides.

Linking-sensitive items Items whose DIF flag changes across the linking assumptions. These are the items to prioritize for review by content and translation experts.

The overall verdict is "robust" when the linkings agree within 'tolerance' logits and the bootstrap keeps at least 'stable_share' of the modes within 'tolerance' of the estimate; otherwise "sensitive".

Usage

```
td_sensitivity(
  dif,
  cut = NULL,
  tolerance = 0.15,
  stable_share = 0.8,
  B = 200,
  n_draws = 100,
  seed = NULL
)
```

Arguments

<code>dif</code>	A 'td_dif' object (from [td_dif()]).
<code>cut</code>	Optional raw-score passing standard for pass-rate impact.
<code>tolerance</code>	Difference in the linking shift (logits) considered substantively negligible.
<code>stable_share</code>	Minimum share of bootstrap modes within 'tolerance' of the estimate for the mode to count as stable.
<code>B</code>	Bootstrap replicates.
<code>n_draws</code>	Posterior draws for each pass-rate impact.
<code>seed</code>	Optional seed.

Details

****What the verdict means.**** It is a statement about **dependence on an untestable assumption**, not an estimate of which linking is correct. In the package's known-truth simulations, "sensitive" verdicts were concentrated in the hardest designs (short tests, small translated-language groups, pervasive DIF), where linking errors are larger for every method; within a given design, the verdict did not separate more accurate estimates from less accurate ones. Disagreement between linkings arose mostly because mean and purified linking are biased under heavy directional DIF. Use the verdict to decide when to report results under several assumptions and to send the linking-sensitive items to content and translation experts; only that review can settle which items should anchor the scale. Items near the flagging boundary can be linking-sensitive even when the verdict is robust.

Value

A 'td_sensitivity' object: '\$linkings' (one row per assumption: 'assumption', 'shift', 'focal_mean', 'se', 'n_flagged', and pass-rate impact columns when 'cut' is given), '\$bootstrap' ('sd', 'lower', 'upper' (90 linking and a 'linking_sensitive' indicator), '\$range' (largest difference between linkings) and '\$verdict'.

Examples

```
sim <- td_simulate(n_ref = 400, n_focal = 120, n_items = 20, seed = 5)
dif <- td_dif(td_calibrate(sim$responses, sim$group))
sens <- td_sensitivity(dif, cut = 12, B = 50, n_draws = 20, seed = 1)
sens
```

td_simulate	<i>Simulate a source-language and a translated administration with known DIF</i>
-------------	--

Description

Items carry binary adaptation features (idiom, cultural referent, measurement units, high vocabulary load). In the translated form an item's difficulty shifts by 'sum(feature_effects * features)' plus small noise, so DIF is directional (translations mostly harder) and unbalanced, which is the case where mean-based linking fails. The focal (translated) group is small and lower-scoring on average.

Usage

```
td_simulate(
  n_ref = 2000,
  n_focal = 150,
  n_items = 60,
  focal_mean = -0.5,
  focal_sd = 1,
  feature_prev = c(idiom = 0.1, cultural = 0.1, units = 0.08, vocabulary = 0.12),
  feature_effects = c(idiom = 0.6, cultural = 0.5, units = -0.4, vocabulary = 0.35),
  dif_noise = 0.1,
  seed = NULL
)
```

Arguments

n_ref, n_focal	Group sizes.
n_items	Test length.
focal_mean, focal_sd	Focal-group ability (reference is N(0, 1)).
feature_prev	Prevalence of each feature.
feature_effects	DIF (logits) contributed by each feature.

dif_noise	SD of feature-unrelated DIF on flagged items.
seed	Optional seed.

Value

An 'td_sim': '\$responses' (0/1 matrix, persons x items), '\$group' ('"ref"/'"focal"'), '\$features' (item data frame), '\$truth' ('b_ref', 'dif', 'focal_mean', 'focal_sd').

Examples

```
sim <- td_simulate(n_ref = 400, n_focal = 120, n_items = 20, seed = 5)
table(dif_item = sim$truth$dif_item)
head(sim$features)
```

Index

td_calibrate, [2](#)
td_dif, [3](#)
td_features, [4](#)
td_impact, [5](#)
td_mh, [6](#)
td_report, [6](#)
td_sensitivity, [7](#)
td_simulate, [9](#)