Package 'aiDIF'

April 21, 2026

Type Package

Title Differential Item Functioning for AI-Scored Assessments

Version 0.1.0

Description Detects and quantifies differential item functioning (DIF) in AI-scored educational and psychological assessments. Provides a fully self-contained robust DIF engine (M-estimation via iteratively re-weighted least squares with the bi-square loss) alongside the novel Differential AI Scoring Bias (DASB) test, which detects item-level scoring shifts that differ across subgroups when comparing human and AI scoring conditions. Includes simulation utilities, anchor weight diagnostics, and an AI-effect classification framework.

License GPL (>= 3)

Encoding UTF-8

Depends R (>= 3.5.0)

Imports Matrix, stats, graphics

Suggests mirt, testthat (>= 3.0.0), knitr, rmarkdown

Config/testthat/edition 3

RoxygenNote 7.3.3

VignetteBuilder knitr

URL https://github.com/causalfragility-lab/aiDIF

BugReports https://github.com/causalfragility-lab/aiDIF/issues

NeedsCompilation no

Author Subir Hait [aut, cre] (ORCID: <https://orcid.org/0009-0004-9871-9677>)

Maintainer Subir Hait <haitsubi@msu.edu>

Repository CRAN

Date/Publication 2026-04-21 20:52:36 UTC

Contents

`ai_effect_summary` *Summarise the effect of AI scoring on DIF flagging.*

Description

Compares the DIF flagging patterns from human and AI scoring conditions and classifies each item as: `"stable_clean"` (not flagged in either), `"stable_dif"` (flagged in both), `"introduced"` (flagged only under AI), `"masked"` (flagged only under human), or `"new_direction"` (flagged in both but bias reverses sign).

Usage

```
ai_effect_summary(dif_human, dif_ai, alpha = 0.05)
```

Arguments

`dif_human`	A `data.frame` returned by `fit_aidif` for the human scoring condition.
`dif_ai`	A `data.frame` returned by `fit_aidif` for the AI scoring condition.
`alpha`	Significance threshold for flagging. Default: `0.05`.

Value

A `data.frame` with one row per item/threshold and columns:

`human_delta` Estimated DIF effect under human scoring.

`ai_delta` Estimated DIF effect under AI scoring.

`human_flag` Logical: flagged under human scoring?

`ai_flag` Logical: flagged under AI scoring?

`status` Classification (see Description).

See Also

scoring_bias_test, fit_aidif

Examples

```
eg <- make_aidif_eg()
mod <- fit_aidif(eg$human, eg$ai)
ai_effect_summary(mod$dif_human, mod$dif_ai)
```

---

anchor_weights *Anchor item weights from the robust AI-DIF procedure.*

Description

Returns the bi-square weights assigned to each item under each scoring condition. Items with weight near zero are effectively excluded from the robust scaling estimate, indicating likely DIF contamination.

Usage

```
anchor_weights(object)
```

Arguments

object	An aidif object from fit_aidif.

Value

A data.frame with columns human_weight and (if AI data were provided) ai_weight. Higher weight means the item is contributing more to the robust scale estimate.

Examples

```
eg <- make_aidif_eg()
mod <- fit_aidif(eg$human, eg$ai)
anchor_weights(mod)
```

`estimate_robust_scale` *Robust DIF scale estimation via IRLS*

Description

Estimates a robust location parameter for the vector of IRT scaling functions using iteratively re-weighted least squares (IRLS) with the bi-square loss. This is the core estimation engine of **aiDIF**.

Usage

```
estimate_robust_scale(
  mle,
  alpha = 0.05,
  scale_by = "pooled",
  tol = 1e-07,
  maxit = 100L
)
```

Arguments

`mle`	A validated mle list.
`alpha`	Significance level controlling the bi-square tuning parameter $k = z_{1-\alpha/2}$. Default 0.05.
`scale_by`	Scaling denominator; passed to `compute_scaling_fn`. Default `"pooled"`.
`tol`	Convergence tolerance. Default `1e-7`.
`maxit`	Maximum IRLS iterations. Default `100`.

Value

A list of class `rdif_fit` with elements:

`est` Estimated robust scale parameter.

`weights` Bi-square item weights.

`rho_value` Value of objective at solution.

`n_iter` Number of iterations used.

`k` Tuning parameter used.

`y` Raw scaling function values.

`vcov_est` Covariance matrix of `y` at solution.

`dif_test` Wald item-level DIF test (data.frame).

`dtf_test` Wald test of differential test functioning.

Examples

```
dat <- simulate_aidif_data(n_items = 5, seed = 1)
fit <- estimate_robust_scale(dat$human)
print(fit$est)
```

`fit_aidif` *Fit the AI-DIF model*

Description

The primary estimation function of **aiDIF**. Runs the robust DIF procedure under both human and AI scoring using the built-in IRLS engine (`estimate_robust_scale`), then tests for Differential AI Scoring Bias (DASB).

Usage

```
fit_aidif(
  human_mle,
  ai_mle = NULL,
  alpha = 0.05,
  scale_by = "pooled",
  tol = 1e-07,
  maxit = 100L
)
```

Arguments

`human_mle`	A validated mle list for human-scored data.
`ai_mle`	A validated mle list for AI-scored data, or `NULL`.
`alpha`	Significance level. Default `0.05`.
`scale_by`	Denominator for standardising intercept differences: `"pooled"` (default), `"ref"`, or `"focal"`.
`tol`	IRLS convergence tolerance. Default `1e-7`.
`maxit`	Maximum IRLS iterations. Default `100`.

Value

An object of class `"aidif"`.

See Also

`estimate_robust_scale`, `scoring_bias_test`, `simulate_aidif_data`

Examples

```
dat <- simulate_aidif_data(n_items = 6, seed = 1)
mod <- fit_aidif(dat$human, dat$ai)
print(mod)
summary(mod)
```

`make_aidif_eg` *Built-in example dataset for aiDIF*

Description

Constructs and returns the built-in example dataset: paired human and AI item parameter estimates for 6 items in two groups, with known DIF and DASB planted at specific items.

Usage

```
make_aidif_eg()
```

Details

The data-generating model includes:

- **Item 1**: DIF under human scoring (intercept +0.5 in focal group).
- **Item 3**: Differential AI Scoring Bias (DASB) — AI scoring adds +0.4 to the focal-group intercept only.
- **Impact**: 0.5 SD (focal group higher on latent trait).
- **AI drift**: uniform +0.1 calibration offset on all items in both groups.

Value

A list with elements `human` and `ai`, each a validated mle list (see `simulate_aidif_data` for format details).

See Also

`simulate_aidif_data`, `fit_aidif`

Examples

```
eg  <- make_aidif_eg()
mod <- fit_aidif(eg$human, eg$ai)
summary(mod)
```

plot.aidif *S3 plot method for class* "aidif".

Description

Produces one of several diagnostic plots depending on type.

Usage

```
## S3 method for class 'aidif'
plot(x, type = "dif_forest", ...)
```

Arguments

x	An object of class "aidif".
type	Character. One of:
	"dif_forest" Forest plot of DIF estimates with 95% confidence intervals for both scoring conditions (default).
	"dasb" Bar chart of DASB estimates with error bars.
	"weights" Dot plot of bi-square anchor weights.
	"rho" Bi-square objective function for human scoring.
...	Additional graphical parameters passed to low-level plot functions.

Value

x, invisibly.

print.aidif *S3 print method for class* "aidif".

Description

Prints a compact summary of the estimated robust scaling parameters and, when available, the number of items flagged for DIF and DASB.

Usage

```
## S3 method for class 'aidif'
print(x, ...)
```

Arguments

x	An object of class "aidif".
...	Further arguments (currently ignored).

Value

x, invisibly.

`read_ai_scored` *Validate and bundle paired human/AI parameter estimates*

Description

Takes two mle lists (one per scoring condition) and returns a validated `aidif_data` object for use in `fit_aidif`.

Usage

```
read_ai_scored(human_mle, ai_mle)
```

Arguments

`human_mle`	An mle list for human-scored data. Must contain `est` (a named list `group.1`, `group.2` of data.frames with columns `a1`, `d1`) and `var.cov` (matching list of covariance matrices).
`ai_mle`	An mle list for AI-scored data in the same format.

Value

A list of class `"aidif_data"` with elements `human` and `ai`.

See Also

`fit_aidif`, `make_aidif_eg`, `simulate_aidif_data`

`scoring_bias_test` *Differential AI Scoring Bias (DASB) test.*

Description

For each item, computes the change in item intercept from human to AI scoring within each group, then tests whether this scoring shift differs significantly across groups. A significant result indicates the AI scoring engine introduces a *group-dependent* parameter distortion — i.e., the AI does not merely re-scale all items uniformly but disfavours (or favours) one group at specific items.

Usage

```
scoring_bias_test(human_mle, ai_mle, fun = "d_fun3")
```

Arguments

human_mle	Output of simulate_aidif_data for human-scored data.
ai_mle	Output of simulate_aidif_data for AI-scored data. Must have the same item/group structure.
fun	Scaling function (passed to the internal scaling function) to use when normalising shifts. Default: "d_fun3".

Details

Estimand. Define the scoring shift in group g for item i threshold j as:

$$\delta_{igj} = d_{igj}^{\text{AI}} - d_{igj}^{\text{Human}}$$

The DASB is $\delta_{i2j} - \delta_{i1j}$. Under $H_0 : \text{DASB}_{ij} = 0$ and independence across scoring conditions and groups,

$$\widehat{\text{Var}}(\text{DASB}_{ij}) = (\sigma_{i1j}^{H})^2 + (\sigma_{i2j}^{H})^2 + (\sigma_{i1j}^{AI})^2 + (\sigma_{i2j}^{AI})^2$$

where each σ^2 is the diagonal element of the corresponding group-specific covariance matrix.

Value

A data.frame with one row per item (per threshold for polytomous items) and columns:

shift_g1 Scoring shift $\delta_{i1} = d_{i1}^{AI} - d_{i1}^{H}$.

shift_g2 Scoring shift $\delta_{i2} = d_{i2}^{AI} - d_{i2}^{H}$.

DASB Differential AI Scoring Bias: $\delta_{i2} - \delta_{i1}$.

se Standard error of DASB under the delta method.

z Wald z-statistic.

p_val Two-tailed p-value.

See Also

fit_aidif, ai_effect_summary

Examples

```
eg <- make_aidif_eg()
scoring_bias_test(eg$human, eg$ai)
```

simulate_aidif_data *Simulate item parameter estimates for the AI-DIF model.*

Description

Generates a synthetic aidif_data-compatible list suitable for benchmarking and method evaluation. The data-generating model contains: classical DIF in the human scoring condition (controlled via dif_items and dif_mag), differential AI scoring bias (controlled via dasb_items and dasb_mag), and a latent group mean difference (impact).

Usage

```
simulate_aidif_data(
  n_items = 10L,
  n_obs = 500L,
  impact = 0.5,
  dif_items = 1L,
  dif_mag = 0.5,
  dasb_items = 3L,
  dasb_mag = 0.4,
  ai_drift = 0.1,
  seed = 42L
)
```

Arguments

n_items	Integer. Number of items. Default: 10.
n_obs	Integer. Approximate number of observations per group, used to scale the covariance matrices. Default: 500.
impact	Numeric. Latent mean difference (group 2 minus group 1) in SD units. Default: 0.5.
dif_items	Integer vector. Indices of items with DIF in the human scoring condition (intercept shift added to group 2). Default: 1.
dif_mag	Numeric. Magnitude of the intercept DIF effect (in IRT metric). Default: 0.5.
dasb_items	Integer vector. Indices of items where AI scoring introduces differential bias (intercept shift added to group 2 in the AI condition only). Default: 3.
dasb_mag	Numeric. Magnitude of the DASB effect. Default: 0.4.
ai_drift	Numeric. Uniform intercept shift applied to ALL items in BOTH groups under AI scoring (simulates AI calibration offset). Default: 0.1.
seed	Integer seed for reproducibility, or NULL. Default: 42.

Details

Rather than simulating item responses and refitting IRT models (which requires additional dependencies), this function directly simulates maximum-likelihood estimates and their asymptotic covariance matrices, consistent with a 2PL model fitted to n_obs observations per group.

Value

A list with elements `human` and `ai`, each formatted identically to the output of `read_ai_scored`. Can be passed directly to `fit_aidif`.

Examples

```
dat <- simulate_aidif_data(
  n_items   = 8,
  n_obs     = 600,
  dif_items = c(1, 2),
  dasb_items = 5
)
mod <- fit_aidif(dat$human, dat$ai)
summary(mod)
```

`summary.aidif` *S3 summary method for class* `"aidif"`.

Description

Prints a detailed report including DIF test tables for each scoring condition, the DASB table, and the AI-effect classification.

Usage

```
## S3 method for class 'aidif'
summary(object, ...)
```

Arguments

`object`	An object of class `"aidif"`.
`...`	Further arguments (currently ignored).

Value

`NULL`, invisibly.

Index