

REPFRAME v1.6

This is a Stata package to calculate, tabulate and visualize Reproducibility and Replicability Indicators. These indicators compare estimates from a multiverse of analysis paths of a robustness analysis — be they reproducibility or replicability analyses — to the original estimate in order to gauge the degree of reproducibility or replicability. The package comes with two commands: `repframe` is the main command, and `repframe_gendata` generates a dataset that is used in the help file of the command to show examples of how the command works.

The package can be installed in Stata by executing:

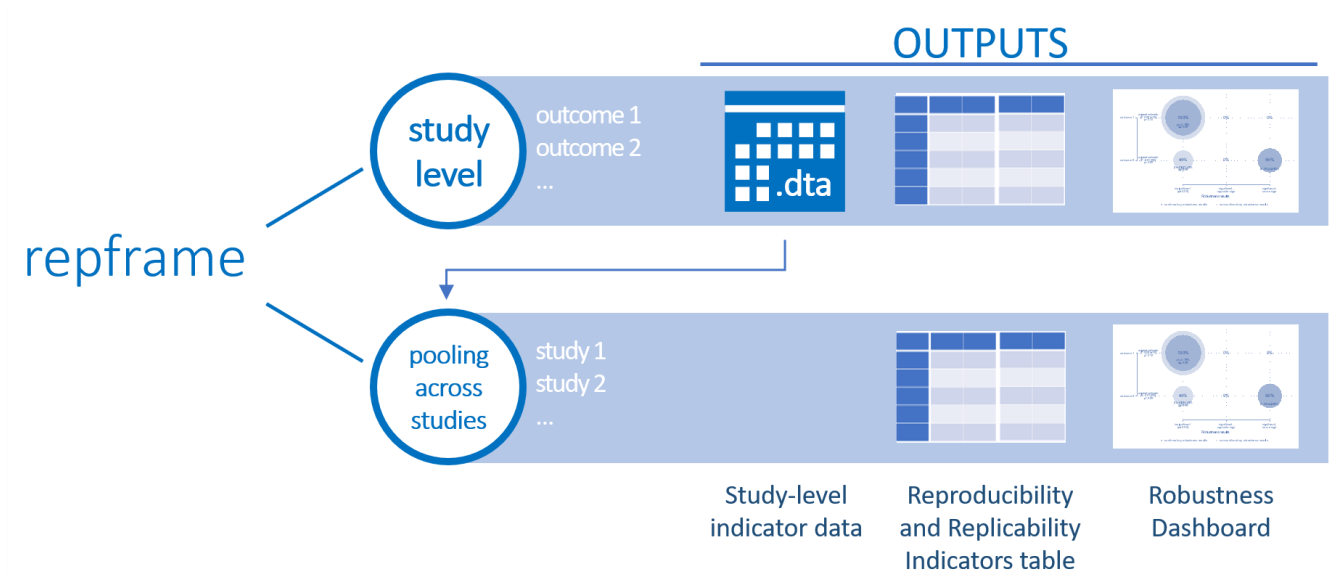
```
net install repframe,
from("https://raw.githubusercontent.com/guntherbensh/repframe/main") replace
```

Once installed, please see

```
help repframe
```

for the syntax and the whole range of options.

As shown in the figure below and described in the following, the `repframe` command can be applied both to derive indicators at the level of individual studies and to pool these indicators across studies. At both levels, the command produces two outputs, a table with the main set of indicators (*Reproducibility and Replicability Indicators table*) as .csv or .xlsx file and a so-called *Robustness Dashboard* that visualizes a second set of indicators. At the study level, the command additionally produces a Stata .dta file as a third output. This *study-level indicator data* is ready to be re-introduced into the command to calculate the indicators across studies.



Defaults applied by the `repframe` command

The **repframe** command applies a few default assumptions. Use the following options in case your analysis makes different assumptions.

Tests for statistical significance: The command applies two-sided t -tests to define which p -values imply statistical significance. These tests may apply different significance levels to the original estimates (**siglevel_orig**(#)) and to the robustness estimates (**siglevel**(#)). If the related p -values retrieved via the option **pval**(varname) are based on one-sided tests, these p -values need to be multiplied by two so as to make them correspond to a two-sided test. If no information on p -values is available, the command derives the missing p -value information applying the t -test formula. Depending on which additional information is available, this may be done based on t / z -scores (**zscore**(varname)), on standard errors (**se**(varname)) and degrees of freedom (**df**(varname)), or on standard errors assuming a normal distribution. Remember that the latter may not always be appropriate, for example with small samples or when estimations have few degrees of freedom because they account for survey sampling, e.g. via the Stata command **svy**:, or when p -values are derived using randomisation inference. Conversely, if input data on p -values is based on other distributional assumptions than normality, the formula may not correctly derive standard errors. It is therefore recommended to specify both the information on p -values and on standard errors, and to consider the implications if non-normality is assumed in either the original or robustness analysis.

[!IMPORTANT]

Replicators of the **Robustness Reproducibility in Economics(R²E)** project should always apply a 5% significance level to the robustness analysis, i.e. **siglevel**(5).

Units in which effect sizes are measured: The command assumes that effect sizes in the original study and in the robustness analysis are measured in the same unit. If this is not the case, for example because one is measured in log terms and the other is not, use the option **sameunits**(varname). This option requires a numerical variable **varname** containing the observation-specific binary information on whether the two are measured in the same unit (1=yes) or not (0=no).

Original analysis to be included as one robustness analysis path: The command assumes that the specification from the original analysis is not to be included as one analysis path in the multiverse robustness analysis. Otherwise specify via the option **orig_in_multiverse**(varname) for which result the specification from the original analysis is supposed to be included in the multiverse robustness analysis. Then, the original specification is incorporated in the computation of three of the **variation indicators** (\$l_{4}\$, \$l_{5}\$, and \$l_{3}\$) for the respective result(s). Irrespective of whether the original specification is included as one robustness analysis path or not, the dataset should include the information on the original specification for each result as a separate analysis path in the input data.

Required input data structure

Data structure for analyses at study level

The input data at study level needs to be in a specific format for **repframe** to be able to calculate the indicators and dashboards. Each observation should represent one analysis path, that is the combination of analytical decisions in the multiverse robustness analysis. In the toy example with one main result (here: one main outcome) represented in the below figure, two alternative choices are assessed for one analytical decision (**analytical_decision_1**, e.g. a certain adjustment of the outcome variable) and three alternative choices are assessed for two other analytical decisions (**analytical_decision_2** and **analytical_decision_3**, e.g.

the set of covariates and the sample used). This gives a multiverse of $3^2 \times 2^1 = 18$ analysis paths, if all combinations are to be considered. The number of observations is therefore 18 in this example.

For each observation, the minimum requirement is that the variable **mainvar** (this is the result at the study level) is defined together with the coefficient information retrieved via the option `beta(varname)` and information to determine statistical significance. The variable **mainvar** should be numeric with value labels. It is recommended to specify both the information on *p*-values and on standard errors, as outlined above in the sub-section on defaults applied by the `repframe` command. As noted in that same sub-section, the dataset should furthermore include, for each result, the same information on the original analysis as one analysis path. It is further recommended to include variables reflecting the *analytical decisions* via the option `decisions()`. Here, each decision variable should be labelled, numeric with the decision adopted by the original authors always being zero, and with labelled values. The `repframe` command gives error messages if these requirements are not met.

	outcome	beta	pval	se	beta_orig	pval_orig	se_orig	analytical_decision_1	analytical_decision_2	analytical_decision_3
1	outcome 1	.3024293	.0136777	.1226586	.1094196	.0128482	.0439798	choice 1	choice 1	choice 1
2	outcome 1	.2595794	.0965241	.1561903	.1094196	.0128482	.0439798	choice 1	choice 1	choice 2
3	outcome 1	.0961636	.0013007	.0299033	.1094196	.0128482	.0439798	choice 1	choice 1	choice 3
4	outcome 1	-.0176124	.6910904	.0443218	.1094196	.0128482	.0439798	choice 1	choice 2	choice 1
5	outcome 1	.0400641	.5888057	.0741148	.1094196	.0128482	.0439798	choice 1	choice 2	choice 2
6	outcome 1	.1750259	.0010347	.0533468	.1094196	.0128482	.0439798	choice 1	choice 2	choice 3
7	outcome 1	.7052199	.2424786	.6033636	.1094196	.0128482	.0439798	choice 1	choice 3	choice 1
8	outcome 1	.3365864	.0797519	.1921016	.1094196	.0128482	.0439798	choice 1	choice 3	choice 2
9	outcome 1	.0957003	.0025346	.0316975	.1094196	.0128482	.0439798	choice 1	choice 3	choice 3
10	outcome 1	.1400244	.0428805	.0691522	.1094196	.0128482	.0439798	choice 2	choice 1	choice 1
11	outcome 1	-.0709329	.9186165	.6942191	.1094196	.0128482	.0439798	choice 2	choice 1	choice 2
12	outcome 1	.6219236	.1261105	.4065867	.1094196	.0128482	.0439798	choice 2	choice 1	choice 3
13	outcome 1	.0066388	.9355342	.0820782	.1094196	.0128482	.0439798	choice 2	choice 2	choice 1
14	outcome 1	.2031607	.1834762	.1527377	.1094196	.0128482	.0439798	choice 2	choice 2	choice 2
15	outcome 1	.6081301	.2814071	.564566	.1094196	.0128482	.0439798	choice 2	choice 2	choice 3
16	outcome 1	.2724592	.0063347	.0998043	.1094196	.0128482	.0439798	choice 2	choice 3	choice 1
17	outcome 1	.215114	.003624	.0739437	.1094196	.0128482	.0439798	choice 2	choice 3	choice 2
18	outcome 1	.1377569	.0000237	.0325898	.1094196	.0128482	.0439798	choice 2	choice 3	choice 3

The Stata help file contains a simple example that uses the command `repframe_gendata` to build such a data structure.

Data structure for analyses across studies

The `repframe` command can also be used to compile Reproducibility and Replicability Indicators across studies. To do so, one only has to append the *study-level indicator data* that include the Reproducibility and Replicability Indicators of individual studies and then feed them back into a variant of the `repframe` command. The following steps need to be taken:

1. run `repframe` multiple times with individual studies to create the *study-level indicator data* saved as `repframe_studydata_[fileidenfier].dta` — with [fileidenfier] as defined by the option `fileidentifier(string)`
2. `append` the individual *study-level indicator data*, making sure that all individual studies applied the same significance level, which can be checked with the variable **siglevel** contained in the *study-level indicator data*
3. run the following commands to compile a dataset with Reproducibility and Replicability Indicators across studies

```
. encode ref, gen(reflist)
. drop ref
. order reflist
. save "[filename].dta", replace
```

where [filename] can be freely chosen for the dataset containing all appended *study-level indicator data*, potentially including the full path of the file.

4. run `repframe` again, now using the option `studypool(1)` to request the calculation of indicators across studies.

The Reproducibility and Replicability Indicators

The *Reproducibility and Replicability Indicators table* and the *Robustness Dashboard* present two separate sets of indicators. These indicators are primarily designed as easily and intuitively interpretable metrics for robustness analyses. First of all, they are applicable to tests on robustness reproducibility, which asks to what extent results in original studies are robust to alternative plausible analytical decisions on the same data (Dreber and Johannesson 2024). This makes it plausible to assume that the tests of robustness reproducibility and the original study measure exactly the same underlying effect size, with no heterogeneity and no difference in statistical power.

For tests of replicability using new data or alternative research designs, more sophisticated indicators are required to account for potential heterogeneity and difference in statistical power (cf. Mathur and VanderWeele 2020, Pawel and Held 2022).

The indicators are meant to inform about the following three pieces of information on reproducibility and replicability, related to either statistical significance or effect sizes:

- agreement indicators: Do the original and robustness analyses exhibit the same level of statistical significance (effect size)?
- relative indicators: To what extent does the statistical significance (do effect sizes) differ between the original and robustness analysis paths?
- variation indicators: To what extent does the statistical significance (do effect sizes) vary across robustness analysis paths?

For situations, in which an original study and a robustness analysis apply different classifications of what constitutes a statistically significant result, i.e. different levels of statistical significance $\{\alpha\}$, the *Robustness Dashboard* adds a

- significance classification agreement indicator: Do the original and robustness analyses agree in terms of whether their findings are classified as statistically significant?

Moreover, the *Robustness Dashboard* additionally includes the option `extended(string)`, which allows incorporating a

- significance switch indicator: To what extent are robustness coefficients (standard errors) large (small) enough to have turned an originally insignificant result significant, regardless of the associated standard error (coefficient)? And what about the reverse for originally significant results?

As one additional feature, the option `ivarweight(1)` allows borrowing inverse-variance weighting from classical meta-analyses in order to account for the noisiness of the different result measures.


Reproducibility and Replicability Indicators table

The following describes the main indicators presented in the *Reproducibility and Replicability Indicators table* as they are computed at the level of each assessed result within a single study. Aggregation across results at the study level is simply done by averaging the indicators as computed at result level, separately for results reported as originally significant and results reported as originally insignificant. Similarly, aggregation across studies is simply done by averaging the indicators as computed at study level. An example of a *Reproducibility and Replicability Indicators table* at study level is provided at the end of this section.

1. The **statistical significance indicator** as a significance agreement indicator measures for each result j the share of the n robustness analysis paths i that are reported as statistically significant or insignificant in both the original study and the robustness analysis. Accordingly, the indicator is computed differently for results where the original estimates were reported as statistically significant and those where the original estimates were found to be statistically insignificant. Statistical significance is defined by a two-sided test with α^{orig} being the significance level applied in the original study and α being the significance level applied in the robustness analysis. For statistically significant original estimates, the effects of the robustness analysis paths must also be in the same direction as the original estimate, as captured by coefficients having the same sign or, expressed mathematically, by $\mathbb{I}(\beta_i \times \beta^{\text{orig}}_j \geq 0)$.

$$I_{\{1\}} = \text{mean}(\mathbb{I}(\text{pval}_i \leq \alpha) \times \mathbb{I}(\beta_i \times \beta^{\text{orig}}_j \geq 0)) \quad \text{if } \text{pval}^{\text{orig}}_j \leq \alpha^{\text{orig}}$$

$$I_{\{1\}} = \text{mean}(\mathbb{I}(\text{pval}_i > \alpha)) \quad \text{if } \text{pval}^{\text{orig}}_j > \alpha^{\text{orig}}$$

 This share indicator is intended to capture whether statistical significance in a robustness analysis confirms statistical significance in an original study. The indicator reflects a combination of *technical* agreement of results (do estimates agree in terms of achieving a certain level of statistical significance?) and *classification* agreement as introduced above (do estimates agree in terms of whether they are classified as statistically significant, given a potentially more or less demanding level of statistical significance applied by original authors?).

Interpretation: An indicator $I_{\{1\}}$ of 0.3 for a result j reported as statistically significant in the original study, for example, implies that 30% of robustness analysis paths for this result (i) are statistically significant according to the significance level adopted in the robustness analysis, and (ii) that their coefficients share the same sign as the coefficient in the original study. Conversely, 70% of robustness analysis paths for this result are most likely statistically insignificant, while it cannot be excluded that part of these paths are statistically significant but in the opposite direction. Note also that robustness analysis paths for this result may be found statistically insignificant — and thus non-confirmatory — only because of a stricter significance level adopted in the robustness analysis compared to the original study. An indicator of 0.3 for results reported as statistically insignificant in the original study implies that 30% of robustness analysis paths for this result are also statistically insignificant according to the significance level adopted in the robustness analysis. Now, the remaining 70% of robustness analysis paths are statistically significant (most likely with the same sign), while a less strict significance level applied in the robustness analysis could now affect this indicator.

2. The **relative effect size indicator** measures the mean of the coefficients β_i of all robustness analysis paths for each result j divided by the original coefficient β^{orig}_j . The indicator requires that effect sizes in the original and robustness analyses are measured in the same units. It is furthermore only applied to results reported as statistically significant in the original study, now — and for the following indicators as well — irrespective of whether in the same direction or not.

$$I_{\{2j\}} = \frac{\text{mean}(\beta_i)}{\beta^{\text{orig}}_j} \quad \text{if } p\text{val}^{\text{orig}}_j \leq \alpha^{\text{orig}}$$

$$I_{\{2j\}} \text{ not applicable} \quad \text{if } p\text{val}^{\text{orig}}_j > \alpha^{\text{orig}}$$

👉 This ratio indicator is intended to capture how the size of robustness coefficients compares to the size of original coefficients.

Interpretation: An indicator $I_{\{2j\}}$ above 1 implies that the mean of the coefficients of all the robustness analysis paths for a statistically significant original result j is — in absolute terms — higher than the original coefficient (while both show in the same direction), with a factor of $I_{\{2j\}}$ (e.g. 1.3). An indicator between 0 and 1 means that the mean coefficient in the robustness analysis paths is lower than the original coefficient (while both show in the same direction), again with a factor of $I_{\{2j\}}$ (e.g. 0.7). An indicator below 0 implies that the two compared statistics have different signs. Here, the absolute value of the mean coefficient in the robustness analysis paths is higher (lower) than the absolute value of the original coefficient if $I_{\{2j\}}$ is above (below) -1.

3. The **relative t/z-value indicator** as a relative significance indicator measures for each result j the mean of the t/z -values ($z\text{score}_i$) of all the robustness analysis paths divided by the t/z -value from the original analysis. The indicator is also only derived for results reported as statistically significant in the original study.

$$I_{\{3j\}} = \frac{\text{mean}(z\text{score}_i)}{z\text{score}^{\text{orig}}_j} \quad \text{if } p\text{val}^{\text{orig}}_j \leq \alpha^{\text{orig}}$$

$$I_{\{3j\}} \text{ not applicable} \quad \text{if } p\text{val}^{\text{orig}}_j > \alpha^{\text{orig}}$$

👉 This ratio indicator is designed to compare the statistical significance in a robustness analysis to that in the original study.

Interpretation: An indicator $I_{\{3j\}}$ above (below) 1 means that the average t/z -value of all robustness analysis paths for result j is — in absolute terms — higher (lower) than the original coefficient, suggesting a higher (lower) level of statistical significance in the robustness analysis. An indicator below 0 additionally implies that the two compared statistics have different signs, where the absolute value of the mean t/z -value in the robustness analysis paths is higher (lower) than the absolute value of the original t/z -value if $I_{\{3j\}}$ is above (below) -1.

4. The **effect size variation indicator** measures for each result j the standard deviation sd of all robustness coefficients divided by the standard error se of the original coefficient. Here, the β_i may incorporate the original specification as one robustness analysis path. The indicator requires that effect sizes of the original and robustness analyses are measured in the same units.

$$I_{\{4j\}} = \frac{sd(\beta_i)}{se(\beta^{\text{orig}}_j)}$$

applied separately to $p\text{val}^{\text{orig}}_j \leq \alpha^{\text{orig}}$ and $p\text{val}^{\text{orig}}_j > \alpha^{\text{orig}}$.

👉 This ratio indicator is intended to capture how the variation in coefficients of a robustness analysis compares to the variation estimated for the original coefficient.

Interpretation: An indicator I_{4j} above (below) 1 means that variation across all robustness analysis paths for result j is higher (lower) than the variation estimated in the original analysis, with a factor of I_{4j} .

5. The **t/z-value variation indicator** as a significance variation indicator measures the standard deviation of t/z-values of all the robustness analysis paths for each result j . Here, the $zscore_i$ may incorporate the original specification as one robustness analysis path.

$$I_{5j} = sd(zscore_i)$$

applied separately to $pval^{orig}_j \leq \alpha^{orig}$ and $pval^{orig}_j > \alpha^{orig}$.

👉 This absolute indicator is intended to capture the variation in the statistical significance across robustness analysis paths.

Interpretation: I_{5j} simply reports the standard deviation of t/z-values of all the robustness analysis paths for result j as a measure of variation in statistical significance. Higher values indicate higher levels of variation.

The following shows an example of the *Reproducibility and Replicability Indicators table*, indicating the five indicators as outlined above. The indicators are grouped by whether the respective result was originally reported as statistically significant or not. Each of these two sets of indicators also includes the average across the respective results.

	A	B	C	D	I_1	I_2	I_3	I_4	I_5	
1	Outcome	Original beta	Original beta	p-value	(RF.1)	(RF.2)	(RF.3)	(RF.4)	E(RF.5)	Sig
2	Outcome 1	0.26	17	0.00	1.00	0.98	1.00	0.35	0.46	
3	Outcome 2	0.11	21	0.01	0.35	-9.18	0.58	769.49	1.12	
4	Outcome 3	11.58	61	0.07	0.61	1.08	1.27	23.94	0.83	
5	All outcomes (originally sig.)			0.03	0.65	-2.37	0.95	264.59	0.80	
6										
7	Outcome 4	-8.35	-26	0.13	0.96			0.56	0.47	
8	All outcomes (originally insig.)			0.13	0.96			0.56	0.47	
9										
10	Notes:									
11	3/4 outcomes originally significant applying original authors' sig. level (10% level)									
12	reframe version 1.4.2 4mar2024									

As additional references, the second to fourth columns of the table contain the following three figures on the estimate from the original study:

- original beta estimate, β^{orig}_j
- original beta estimate, expressed as % deviation from original mean of the outcome, i.e. $\beta^{orig}_j / \text{mean}^{orig}_j \times 100$
- p-value of original beta estimate, $pval^{orig}_j$.

A key feature of the dashboard is that indicators are tailored to specific sub-groups of analysis paths. Both for the original study and for the robustness analysis, it distinguishes between results that are statistically significant and those that are not, and by whether results that are significant in the robustness analysis have the same or an opposite sign as the result in the original study. For example, the *relative effect size indicator* is only derived for robustness analysis paths that are statistically significant and in the same direction as the original estimate. The idea is to restrict the indicator to those analysis paths, for which it is most meaningful (instead of averaging the indicator across all analysis paths for the respective result), and limit the information for the other analysis paths — those that are statistically significant or statistically significant but in the opposite direction — to a more simple and parsimonious set of indicators.

The dashboard includes up to nine indicators. The core set is composed of four default indicators, I'_1 to I'_4 , with two conditional indicators, I'_5 and I'_6 , and an extended version of the dashboard additionally includes indicators I'_7 to I'_9 .

In the same vein as for indicators presented in the *Reproducibility and Replicability Indicators table*, aggregation across results at the study level (across studies) is simply done by averaging the indicators as computed at result (study) level, separately for originally significant and originally insignificant results. When aggregating across results, a general difference with the indicators included in the *Reproducibility and Replicability Indicators table* becomes the more relevant: the *Robustness Dashboard* applies the same level of statistical significance to original and robustness analyses in classifying results into significant versus insignificant. The motivation is to separate *technical* and *classification* agreement of results as defined above and outlined in the description of the first two indicators.

An example of a *Robustness Dashboard* at study level is provided at the end of this section.

1. The **significance agreement indicator** is derived for each result j in a similar way as the *statistical significance indicator* from the *Reproducibility and Replicability Indicators table*. The only differences are that (i) the indicator is the same for statistically significant and insignificant estimates in robustness analyses and that (ii) the same significance level α is applied to the original analysis and to the robustness analysis (note that (ii) only becomes relevant when aggregating into significant versus insignificant original results, as done in the [bottom Dashboard figure below](#)). The indicator is expressed in % of all robustness analysis paths on either statistically significant or insignificant original results and, hence, additionally multiplied by 100. For statistically significant robustness analysis paths with same sign, the indicator is calculated as follows:

$$I'_1 = \text{mean}(\mathbb{I}(\text{pval}_i \leq \alpha) \times \mathbb{I}(\beta_i \times \beta^{\text{orig}}_j \geq 0)) \times 100$$

applied separately to $\text{pval}^{\text{orig}}_j \leq \alpha$ and $\text{pval}^{\text{orig}}_j > \alpha$. The same indicator is also calculated for statistically significant robustness analysis paths with opposite sign, i.e. differing from the above formula through $\mathbb{I}(\beta_i \times \beta^{\text{orig}}_j < 0)$. For statistically insignificant robustness analysis paths, the indicator corresponds to 100 minus these two indicators on statistically significant results with same and opposite sign.

👉 This proportion indicator is intended to capture the *technical* agreement of results (are estimates robust in terms of achieving a certain level of statistical significance?).

Interpretation: An indicator I'_1 of 30% implies that 30% of robustness analysis paths for result j are statistically significant. Depending on which of the four sub-indicators of the *Robustness Dashboard*

one is referring to, this refers to (i) statistically significant *or* insignificant original results and to (ii) original and robustness coefficients that share *or* do not share the same sign. For example, if I'_{1j} is 30% for results with the same sign and 3% for results with opposite signs, the remaining 67% of robustness analysis paths for this result are statistically insignificant. The significance levels applied to the original study and the robustness analysis are identical and correspond to the one defined in the robustness analysis.

2. The **relative effect size indicator** differs from I'_{2j} from the *Reproducibility and Replicability Indicators table* in that it is only derived for robustness analysis paths that are (i) statistically significant and (ii) in the same direction as the original estimate. In addition, the indicator takes the median of the robustness coefficients instead of the mean, in order to be less sensitive to outliers. Furthermore, one is subtracted from the ratio, in order to underscore the relative nature of the indicator. A ratio of 2/5 thus turns into -3/5, and multiplied by 100 to -60%.

$$I'_{2j} = \left(\frac{\text{median}(\beta_i)}{\beta^{\text{orig}}_j} - 1 \right) \times 100 \quad \text{if } p\text{val}^{\text{orig}}_j \leq \alpha \text{ and } p\text{val}_i \leq \alpha \text{ and } \beta_i \times \beta^{\text{orig}}_j \geq 0$$

$$I'_{2j} \text{ not applicable otherwise}$$

👉 This ratio indicator is intended to capture how effect sizes of robustness analyses compare to the original effect sizes. The indicator focuses on the case where a comparison of effect sizes is most relevant and interpretable, that is when both the original and robustness analysis yield estimates that are statistically significant and in the same direction.

Interpretation: An indicator I'_{2j} for result j with an originally significant result (below) 0% means that the mean of statistically significant robustness coefficients in the same direction as the original estimate is higher (lower) than the original coefficient, by $I'_{2j}\%$ — e.g. +30% (-30%).

The *Robustness Dashboard* does not include a **relative significance indicator**.

3. The **effect size variation indicator** measures the mean absolute deviation of coefficients in robustness analysis paths from their median. Like I'_{2j} , it only considers robustness analysis paths for results reported as statistically significant that are (i) statistically significant and (ii) in the same direction as the original estimate. The mean value is divided by the original coefficient and multiplied by 100 so that it is measured in the same unit as I'_{2j} . Here, the β_i may incorporate the original specification as one robustness analysis path. The indicator requires that effect sizes in the original and robustness analyses are measured in the same units.

$$I'_{3j} = \frac{\text{mean}(|\beta_i - \text{median}(\beta_i)|)}{\beta^{\text{orig}}_j} \times 100 \quad \text{if } p\text{val}^{\text{orig}}_j \leq \alpha \text{ and } p\text{val}_i \leq \alpha \text{ and } \beta_i \times \beta^{\text{orig}}_j \geq 0$$

$$I'_{3j} \text{ not applicable otherwise}$$

👉 This ratio indicator is intended to capture how the variation in coefficients of robustness analysis paths compares to the variation in original coefficients. The indicator complements I'_{2j} focusing on the case of original and robustness analyses with estimates that are statistically significant and in the same direction.

Interpretation: An indicator I'_{3j} of, for example, 10% means that variation across robustness analysis paths for result j is equivalent to 10% of the original coefficient.

1. The **significance variation indicator** measures the mean of the deviations between p -values from the robustness analysis paths and the original p -value. This indicator is always derived, except for robustness and original analyses with both statistically significant estimates, since the deviation is known to be small in that case.

$$I'_{4j} = \text{mean}(\mid pval_j - pval^{\text{orig}}_j \mid)$$

applied separately to (i) $pval^{\text{orig}}_j \leq \alpha \text{ \& \& } pval_j > \alpha$, (ii) $pval^{\text{orig}}_j > \alpha \text{ \& \& } pval_j > \alpha$, and (iii) $pval^{\text{orig}}_j > \alpha \text{ \& \& } pval_j \leq \alpha$.

👉 This absolute indicator is intended to capture the variation in statistical significance across robustness analysis paths that turned or are statistically insignificant.

Interpretation: An indicator I'_{4j} of 0.2, for example, implies that p -values among certain robustness analysis paths for result j on average differ by 0.2 from the original p -value. Depending on which of the three sub-indicators of the *Robustness Dashboard* one is referring to, this refers to the case of (i) a significant estimate in the original analysis and insignificant estimates in the robustness analysis, (ii) an insignificant estimate in the original analysis and insignificant estimates in the robustness analysis, or (iii) an insignificant estimate in the original analysis and significant estimates in the robustness analysis. Like p -values themselves, this deviation may assume values between 0 (very small deviation) and 1 (maximum deviation).

5. The **effect size agreement indicator** measures the share of robustness coefficients that lie inside the bounds of the confidence interval of the original coefficient when applying the significance level α adopted in the robustness analysis, $\beta(\text{cilo})^{\text{orig}}_j$ and $\beta(\text{ciup})^{\text{orig}}_j$. It only considers statistically insignificant robustness analysis paths for results reported as statistically significant in the original study. The indicator requires that effect sizes in the original and robustness analyses are measured in the same units.

$$I'_{5j} = \text{mean}(\mathbb{I}(\beta(\text{cilo})^{\text{orig}}_j \leq \beta_j \leq \beta(\text{ciup})^{\text{orig}}_j)) \times 100 \quad \text{if } pval^{\text{orig}}_j \leq \alpha \text{ \& \& } pval_j > \alpha$$

$$I'_{5j} \text{ \text{not applicable otherwise}}$$

👉 This proportion indicator is intended to complement the *significance agreement indicator* and thereby to capture *technical* agreement of results not only in terms of achieving a certain but arbitrary level of statistical significance, but also in terms of showing similarity of coefficients.

Interpretation: An indicator I'_{5j} of 10% implies that 10% of robustness analysis paths for this result j with originally significant results are insignificant according to the significance level adopted by the robustness analysis, but with robustness coefficients that cannot be rejected to lie inside the confidence interval of the estimate. The closer these 10% are to the share of statistically insignificant robustness analysis paths for this result, the less does this indicator confirm the *statistical significance indicator*. For example, if the share of statistically insignificant robustness analysis paths for this result is 15%, two-thirds of these analysis paths are non-confirmatory according to *statistical significance indicator* and confirmatory according to the *effect size agreement indicator*.

6. The **indicator on non-agreement due to significance classification** is an indicator that focuses on *classification* robustness of results as defined above. It applies only in situations in which an original study applied a different — more or less stringent — classification of what constitutes a statistically

significant result than the robustness analysis. Specifically, it identifies those originally significant (insignificant) results that have statistically insignificant (significant) robustness analysis paths only because a more (less) stringent significance level definition is applied in the robustness analysis than in the original study. The indicator is also expressed in % and therefore includes the multiplication by 100. For the case where a more stringent significance level definition is applied in the robustness analysis, the indicator is calculated as follows.

$$I'_{\{6j\}} = \text{mean}(\mathbb{I}(\alpha < pval_i \leq \alpha^{\{orig\}})) \times 100 \quad \text{if } \alpha < pval^{\{orig\}}_j \leq \alpha^{\{orig\}}$$

$$I'_{\{6j\}} \text{ not applicable otherwise}$$

In the opposite case, with a less stringent significance level definition applied in the robustness analysis, the same formula applies with opposite signs.

👉 This proportion indicator is intended to capture non-robustness of findings reported as (in)significant in original studies that is due to differences in the classification of statistical significance.

Interpretation: Consider the case where the robustness analysis paths apply a significance level of 5% and the original analysis applied a less strict significance level of 10%. In this case, estimates from robustness analyses with $0.05 < pval_i \leq 0.10$ are only categorized as insignificant and thus having a non-agreeing significance level because of differing definitions of statistical significance. An indicator $I'_{\{6j\}}$ of 10%, for example, implies that this holds true for 10% of robustness analysis paths for result j .

7. The **significance classification agreement indicator** aggregates the information on significance classification agreement between the robustness analysis and the original analysis.

$$I'_{\{7j\}} = I'_{\{1j\}}^{\{ssign\}} \times 100 \quad \text{if } pval^{\{orig\}}_j \leq \alpha^{\{orig\}}$$

$$I'_{\{7j\}} = (1 - I'_{\{1j\}}^{\{ssign\}} - I'_{\{1j\}}^{\{nsign\}}) \times 100 \quad \text{if } pval^{\{orig\}}_j > \alpha^{\{orig\}}$$

where *ssign* refers to $I'_{\{1j\}}$ when derived for estimates from robustness analyses with the same sign, and *nsign* when derived for estimates from robustness analyses with the opposite sign.

The indicator presented in the *Robustness Dashboard* is the average across all results or studies.

$$I'_{\{7\}} = \text{mean}(I'_{\{7j\}})$$

This is different from the other indicators, as it is not differentiated by whether results are originally significant or insignificant.

In cases where the robustness analysis and original study or studies applied different significance levels, the *Robustness Dashboard* additionally shows this indicator when applying a uniform significance level, that is when the formulae include α instead of $\alpha^{\{orig\}}$. Both indicators have their advantages and disadvantages. Consider the example with $pval^{\{orig\}}_j = 0.07$, $\alpha = 0.05$, and $\alpha^{\{orig\}} = 0.10$. Here, the former indicator would categorize robustness analysis paths with equal p -values of $pval_i = 0.07$ as non-confirmatory, whereas the latter indicator would categorize robustness analysis paths with lower p -values of $pval_i = 0.04$ as non-confirmatory, both of which can be seen as contrary to common intuition. It is therefore generally recommended to use the same significance level in a robustness analysis as in the original study or studies (if the latter differ among each other, the less stringent significance level is to be chosen).

👉 This proportion indicator is intended to capture to which degree statistical significance as reported in original studies is confirmed through the robustness analyses — where the classification of statistical significance may differ from that of the original study or not.

Interpretation: An indicator I'_{7j} of 80% implies that the classification into significant or insignificant in robustness analysis paths confirms the classification by original authors in 80% when averaged over individual results (studies).

Extension of the Robustness Dashboard

The *Robustness Dashboard* additionally includes the option `extended(string)` to show the following type of indicators in an extended set of indicators.

8. & 9. The **significance switch indicators** include two sub-indicators for originally significant and insignificant results, respectively. For originally significant results, these indicators measure the share of robustness coefficients (standard errors) that are sufficiently small (large) to have turned the result insignificant when standard errors (coefficients) are held at their values in the original study. Whether absolute values of coefficients (standard errors) are sufficiently small (large) is determined based on the threshold values $\beta(\text{tonsig})_j$ and $\text{se}(\text{tonsig})_j$. The indicators require that effect sizes in the original and robustness analyses are measured in the same units.

$$I'_{8j} = \text{mean}(\mathbb{I}(\mid \beta_i \mid \leq \beta(\text{tonsig})_j)) \times 100 \quad \text{if } pval^{\text{orig}}_j \leq \alpha \text{ and } pval_i > \alpha$$

$$I'_{9j} = \text{mean}(\mathbb{I}(\text{se}_i \geq \text{se}(\text{tonsig})_j)) \times 100 \quad \text{if } pval^{\text{orig}}_j \leq \alpha \text{ and } pval_i > \alpha$$

The indicators for originally insignificant results are a mirror image of those for originally significant results: now the indicators measure the shares of robustness coefficients (standard errors) that are sufficiently large (small) to have turned results significant, applying threshold values $\beta(\text{tosig})_j$ and $\text{se}(\text{tosig})_j$, respectively.

$$I'_{8j} = \text{mean}(\mathbb{I}(\mid \beta_i \mid > \beta(\text{tosig})_j)) \times 100 \quad \text{if } pval^{\text{orig}}_j > \alpha \text{ and } pval_i \leq \alpha$$

$$I'_{9j} = \text{mean}(\mathbb{I}(\text{se}_i < \text{se}(\text{tosig})_j)) \times 100 \quad \text{if } pval^{\text{orig}}_j > \alpha \text{ and } pval_i \leq \alpha$$

👉 These proportion indicators are intended to capture the drivers behind changes in statistical significance between original study and robustness analysis.

Interpretation: An indicator I'_{8j} of, for example, 30% for a result j with an originally significant result, implies that 30% of the robustness analysis paths that are statistically insignificant have coefficients that are sufficiently small for the robustness analysis path to be statistically insignificant even if the standard error would be identical to the one in the original study. The other (sub-)indicators can be interpreted analogously.

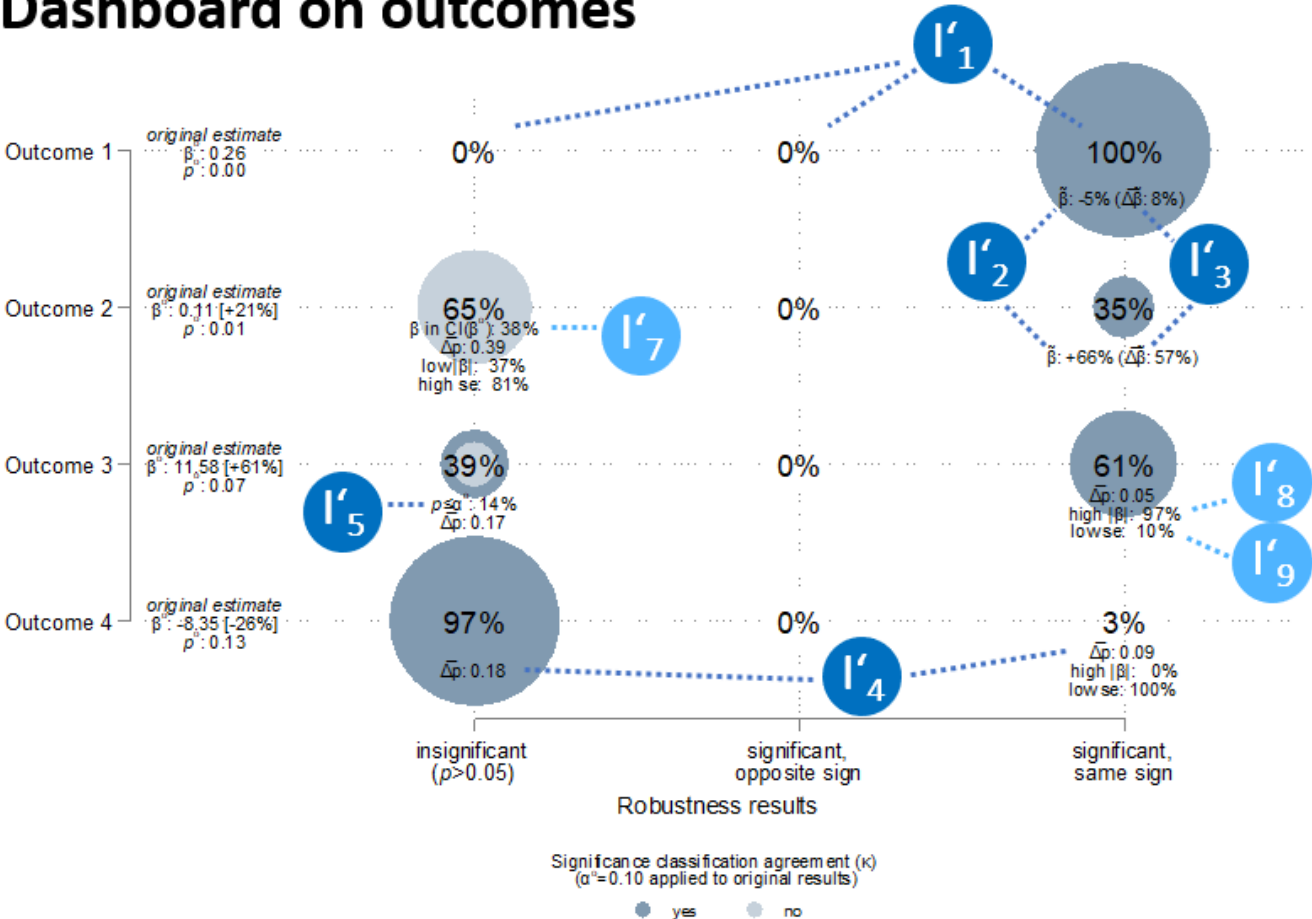
The Dashboard output

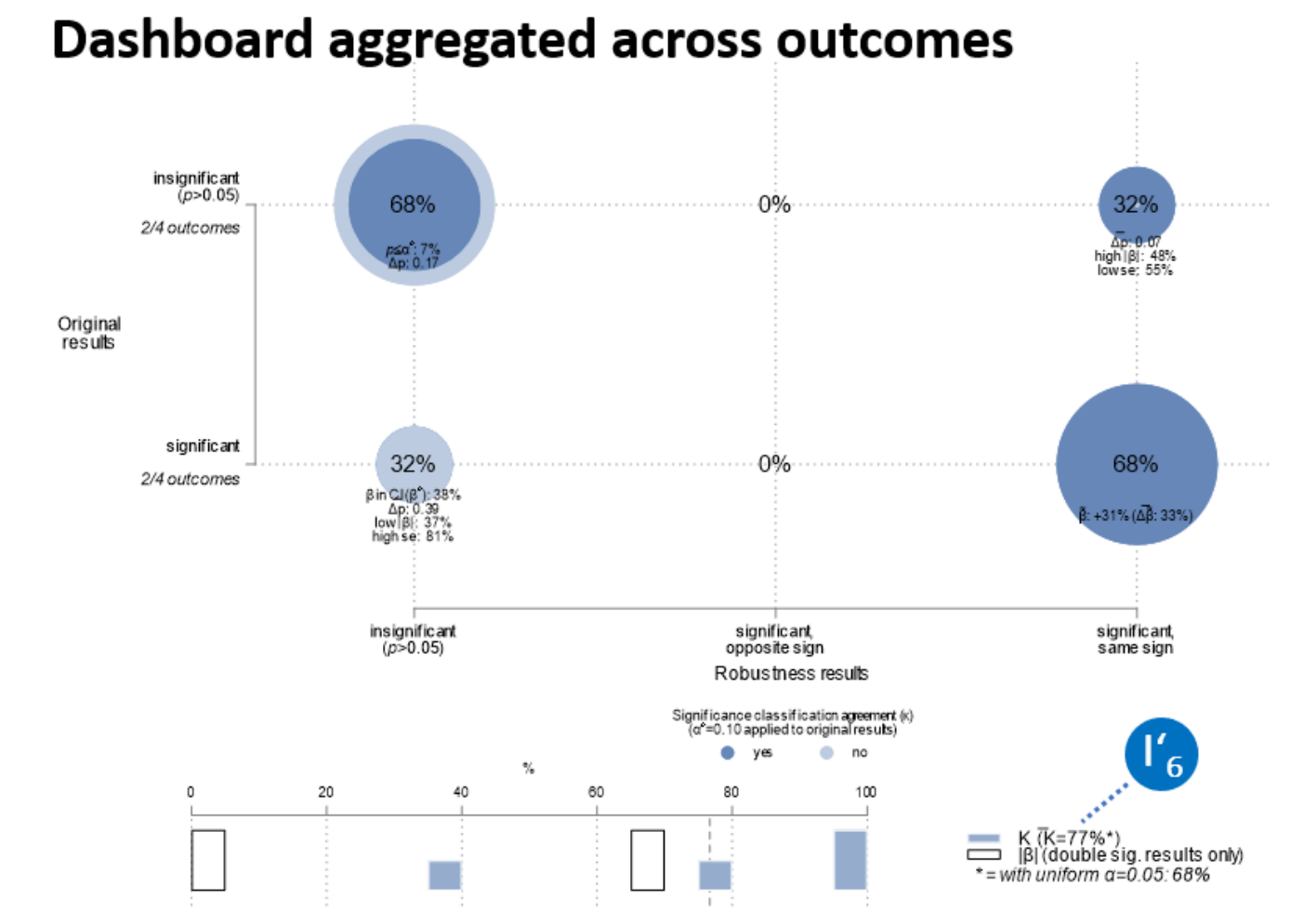
The following shows an example of the *Robustness Dashboard*, indicating where the indicators outlined above can be found in the figure. Indicators from the extended set are in lighter blue. The vertical axis of the dashboard shows individual results, grouped into statistically significant and insignificant if aggregated. Note that this grouping may differ from the one in the *Reproducibility and Replicability Indicators table*, because that table applies to original results the significance level defined by original authors, whereas the dashboard applies the same significance level as adopted in the robustness analysis. The horizontal axis distinguishes between analysis paths with statistically significant and insignificant estimates, additionally differentiating between statistically significant estimates in the same and in opposite direction as the estimates from the original study. Circle sizes illustrate $I'_{\{j\}}$, the *significance agreement indicator*. They are coloured in either darker blue for confirmatory results or in lighter blue for non-confirmatory results. As can be seen with Outcome 3 in the figure, this colouring also discriminates $I'_{\{1j\}}$ from $I'_{\{6j\}}$, the *indicator on non-agreement due to significance classification*.

When aggregating across results or studies, the bottom of the dashboard additionally includes a histogram with the share of confirmatory results and absolute values of effect sizes.

In the results window of Stata, the `repframe` command provides as additional information the (minimum and maximum) number specifications that have been used to derive the dashboard indicators.

Dashboard on outcomes





Summary

The following table summarizes which indicators are included in the *Reproducibility and Replicability Indicators table* and the *Robustness Dashboard*.

Type of indicator		Reproducibility and Replicability Indicators table	Robustness Dashboard	Symbol in Dashboard
significance (sig.)	sig. agreement	$I_{\{1\}}$	$I'_{\{1\}}$	(main figure in dashboard - no symbol)
	relative sig.	$I_{\{3\}}$	-	-
	sig. variation	$I_{\{5\}}$	$I'_{\{4\}}$	$\overline{\Delta p}$ (mean abs. var. of p -value)
	sig. classification agreement	-	$I'_{\{6\}}$ (if different sig. levels)	$p \leq \alpha$ (less stringent sig. level applied in original study) or $p > \alpha$ (more stringent sig. level applied in original study)

Type of indicator		Reproducibility and Replicability Indicators table	Robustness Dashboard	Symbol in Dashboard
	overall sig. (and sig. classification) agreement	-	I'_{7} (if aggregated)	$\overline{\kappa}$ (mean share of confirmatory results)
	sig. switch	-	I'_{8} & I'_{9} (ext.)	high/ low β (abs. value of β) and se
	effect size (e.s.)	-	I'_{5}	β in $CI(\beta^o)$ (confidence interval of orig. β)
	relative e.s.	I_{2}	I'_{2}	$\widetilde{\beta}$ (median β)
	e.s. variation	I_{4}	I'_{3}	$\overline{\Delta\beta}$ (mean abs. var. of β)

Update log

2024-0X-XX, v1.6:

- Include option `decisions()` that allows accounting for the analytical decisions taken.
- Because of that inclusion, the input data structure needed to be revised in that the specifications adopted by the original authors are included as individual analysis paths, and not only via the variables ending with `_orig`; this allows for a straightforward way of specifying the `decisions()` taken by the original authors for each of the results.
 - in the process of this revision, the option `origpath()` was introduced to specify the original authors' specifications, one for each result.
 - relatedly, all `_orig` options — except `siglevel_orig` — have been removed from the `repframe` command.
- The option `orig_in_multiverse()` now requests a variable instead of a simple 0-1 indicator; this allows for setting this option differently for individual results.
- Resolve [Issue #1](#), that is providing an explanation of `beta_rel_orig` in the github Readme.
- Data at level of analysis path with uniform set of variables and uniform naming stored as `repframe_analysispathdata_[fileidenfier].dta`.
- Study-level data stored as `repframe_studydata_[fileidenfier].dta` instead of `repframe_data_[fileidenfier].dta`
- Adjustments affecting the Robustness Dashboard:
 - inclusion of *effect size agreement indicator* into the default set of indicators; accordingly, the option `extended()` can now only be set to "none" or "SIGswitch".
 - split longer results names into two lines.
 - minor adjustment in line spacing of Dashboard for newer Stata versions (version ≥ 16).
- Text revisions, among others regarding the use of the term "result".

2024-06-03, v1.5.2:

- Minor adjustments in Sensitivity Dashboard:
 - rename dashboard to *Robustness Dashboard*, including the option `sensdash()`, which is now called `dashboard()`.
 - correct calculation of the *effect size variation indicator* when `sameunits(variable==0)` and `orig_in_multiverse(1)` applies for any analysis path.
 - correct calculation of the *indicator on non-agreement due to significance classification* when `aggregation(1)`.
 - remove slight inconsistency in rounding if sum of shares would exceed or fall below 100%.
 - adjust colouring of confirmatory and non-confirmatory results.
 - extend *indicator on non-agreement due to significance classification* to situations in which an original study applied a less stringent classification of what constitutes a statistically significant result than the robustness analysis.
 - show histogram with share of confirmatory results (κ) and absolute values of effect sizes (β) at bottom of dashboard when `aggregation(1)`.
 - inclusion of the *overall significance (and sig. classification) agreement indicator*.

2024-03-17, v1.5.1:

- Improve MacOS compatibility.
- Revise table output, including added `tabfmt()` option.
- Fix a bug that occurred when assessing only one result.

2024-03-05, v1.5:

- Fix minor bugs occurring with `studypooling(1)`.
- Clarify that `repframe` only works with Stata 14.0 or higher.
- Introduce option `studypooling()` to `repframe_genadd` command and include illustrative example on pooling across studies in Stata help file.

2024-03-04, v1.4.2:

- Adjust the option `extended()` to allow for multiple choices.
- Extend the application of the *significance variation indicator* in the Sensitivity Dashboard to originally insignificant results with significant robustness analysis paths.
- Add examples to the Stata help file.
- Minor revisions of the code.

2024-02-29, v1.4.1:

- Make options `siglevel()` and `siglevel_orig()` compulsory for analyses at study level.
- Add recommendation to include both the information on *p*-values and standard errors at study level.

2024-02-28, v1.4:

- Add option `siglevel_orig()` to allow testing against significance level adopted by original authors; incorporated as an indicator on significance classification into the Sensitivity Dashboard.
- Additional effect size agreement / confidence interval coverage indicator and additional notes to Sensitivity Dashboard and Reproducibility and Replicability Indicators table.

- Produce Reproducibility and Replicability Indicators table for indicators pooled across studies.
- Remove certain requirements to the input data formatting.
- Use NHANES II data for the example in the help file, among others to have multiple results that effectively differ from each other.
- Revise entire command structure and adopt uniform naming convention.

2024-02-13, v1.3.1:

- Minor amendments to the code.

2024-01-22, v1.3:

- Add the option `studypooling()` to calculate indicators across studies.

2024-01-19, v1.2:

- Incorporate the package `sensdash()`.

2024-01-18, v1.1:

- First version of `repframe` package.

References

Dreber, A. & Johanneson, M. (2024). A Framework for Evaluating Reproducibility and Replicability in Economics. *Economic Inquiry*. doi: [10.1111/ecin.13244](https://doi.org/10.1111/ecin.13244).

Mathur, M. B., & VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(3), 1145-1166.

Pawel, S., & Held, L. (2022). The sceptical Bayes factor for the assessment of replication success. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3), 879-911.