Online Appendix

Do decisions about outliers and influential effects matter?
Evidence from 358 behavioral science meta-analyses

Tomas Havranek[a,b,c], Zuzana Irsova[a,c], Martina Luskova[a], and T. D. Stanley[d]

[a]Charles University, Prague
[b]Centre for Economic Policy Research, London
[c]Meta-Research Innovation Center at Stanford
[d]Deakin University

This online appendix reports, first, the pre-registered summaries not reproduced in the main text (the per-field breakdown of the three outcomes, the operation counts, the baseline levels, an edge-case ledger, and the two registered distribution figures), and then the full verdict cross-tabulations, the weight and estimator diagnostics, the worked example, and the threshold grid behind the results in the article.

Materials. The study was pre-registered on the Open Science Framework (https://doi.org/10.17605/OSF.IO/97CMV). The replication package, with the data and the analysis code, is on Zenodo (https://doi.org/10.5281/zenodo.21216506). The article and this appendix are available at https://meta-analysis.cz/outliers.

A.1 Pre-registered summaries reported in full

A.2 Verdict cross-tabulations in full

A.3 Exploratory detail

A.4 Threshold grid

Table A.1. *The pre-registered per-field breakdown of the three outcomes; the fields differ in dataset and effect metric as much as in field, so the differences are read descriptively.*

Field	Rule	Median $\lvert\Delta d\rvert$	Sig. switch	SESOI-0.20 switch
Psychology ($n = 259$; RE)				
	Drop-extreme	0.0308	1.93%	3.86%
	Studentized	0.0000	1.55%	2.71%
	Winsorize	0.0065	1.94%	0.78%
	DFBETAS	0.0403	3.47%	5.02%
Psychology ($n = 259$; UWLS)				
	Drop-extreme	0.0149	1.54%	4.25%
	Studentized	0.0166	2.32%	8.11%
	Winsorize	0.0043	1.16%	2.32%
	DFBETAS	0.0433	4.25%	7.72%
Psychotherapy ($n = 20$; RE)				
	Drop-extreme	0.0110	0.00%	5.00%
	Studentized	0.0265	0.00%	0.00%
	Winsorize	0.0202	0.00%	0.00%
	DFBETAS	0.0821	0.00%	5.00%
Psychotherapy ($n = 20$; UWLS)				
	Drop-extreme	0.0045	0.00%	0.00%
	Studentized	0.0088	0.00%	0.00%
	Winsorize	0.0045	0.00%	0.00%
	DFBETAS	0.0277	0.00%	0.00%
Exercise ($n = 79$; RE)				
	Drop-extreme	0.0551	11.39%	16.46%
	Studentized	0.0000	7.59%	10.13%
	Winsorize	0.0149	5.06%	5.06%
	DFBETAS	0.0681	8.86%	12.66%
Exercise ($n = 79$; UWLS)				
	Drop-extreme	0.0331	7.59%	8.86%
	Studentized	0.0000	5.06%	6.33%
	Winsorize	0.0051	6.33%	1.27%
	DFBETAS	0.0448	15.19%	8.86%

Note: This is the pre-registered per-field breakdown, computed separately within each field. Descriptive only: field is fully confounded with dataset and effect-size metric, so between-field differences cannot be attributed to field alone, and psychotherapy has only $n = 20$ meta-analyses. The unit is the meta-analysis-by-estimator cell. "Median $\lvert\Delta d\rvert$" is the median absolute change in the pooled effect (Cohen's d) relative to the untreated baseline. "Sig. switch" is the share of cells whose statistical-significance verdict (significant iff two-sided $p < 0.05$) switches under the rule; "SESOI-0.20 switch" is the share whose smallest-effect-size-of-interest verdict ($\lvert d\rvert \geq 0.20$) switches. The four active rules are drop-extreme (delete the single registered most-extreme estimate), studentized ($\lvert$externally studentized deleted residual$\rvert > 3$), winsorize (cap effects beyond the within-meta-analysis 5th/95th quantiles), and DFBETAS (flag $\lvert\text{DFBETAS}\rvert > 2/\sqrt{k}$). Under RE, the studentized and winsorize psychology cells are computed on $n = 258$ (one meta-analysis, Sladekova/A181_5, is not estimable there; see Table A.4). RE = random-effects; UWLS = unrestricted weighted least squares.

Table A.2. *Drop-extreme always removes exactly one estimate; the studentized rule changes nothing in nearly half of all cells; winsorize almost always acts but on few estimates.*

Rule	Cells	Changes nothing	Estimates acted on: median (IQR width; max)
Drop-extreme	716	0.0%	1 (0; 1)
Studentized	715	48.4%	1 (0; 18)
Winsorize	715	0.3%	2 (2; 90)
DFBETAS	716	4.9%	2 (2; 55)

Note: The unit is the meta-analysis-by-estimator cell, pooled over both estimators (RE and UWLS) and all fields. "Cells" is the number of estimable active cells for the rule; studentized and winsorize are computed on 715 rather than 716 because one RE cell (Sladekova/A181_5) is not estimable under those two rules (see Table A.4). "Changes nothing" is the share of cells in which the rule removes or caps no estimate (registered count $n_{changed} = 0$). "Estimates acted on" is, among the acting cells only ($n_{changed} > 0$), the median, the width of the interquartile range, and the maximum number of estimates the rule removes (drop-extreme, studentized, DFBETAS) or caps (winsorize). Drop-extreme deletes the single registered most-extreme estimate by construction, so it always acts on exactly one; the studentized rule ($|$externally studentized deleted residual$| > 3$) is inert in 48.4% of cells and, when it acts, usually removes one estimate but up to 18; winsorize caps a median of two estimates (5th/95th within-meta-analysis quantiles), up to 90 in a large meta-analysis; DFBETAS ($|\text{DFBETAS}| > 2/\sqrt{k}$) is inert in 4.9% of cells. RE = random-effects; UWLS = unrestricted weighted least squares.

Table A.3. *At baseline, before any outlier handling, most meta-analyses are already significant and reach the SESOI-0.20 threshold; psychotherapy is significant in every case.*

Field	Estimator	n	Share sig.	Share SESOI-0.20	Median $\lvert d \rvert$
Psychology	RE	259	85.3%	79.5%	0.421
	UWLS	259	87.3%	78.4%	0.378
Psychotherapy	RE	20	100.0%	90.0%	0.494
	UWLS	20	100.0%	85.0%	0.355
Exercise	RE	79	70.9%	69.6%	0.285
	UWLS	79	60.8%	54.4%	0.219
All	**All**	**716**	**82.5%**	**75.7%**	**0.359**

Note: Baseline is the untreated (no-outlier-handling) meta-analysis, treatment "none"; every baseline cell is estimable, so n is the full count per field per estimator. The unit is the meta-analysis-by-estimator cell. "Share sig." is the share of baseline cells that are statistically significant (two-sided $p < 0.05$); "Share SESOI-0.20" is the share reaching the smallest-effect-size-of-interest threshold ($|d| \geq 0.20$); "Median $|d|$" is the median baseline absolute pooled effect (Cohen's d). The final row pools all 716 baseline cells: 82.5% are significant (591/716), 75.7% reach SESOI-0.20, and the median baseline $|d|$ is 0.359. Because so many meta-analyses start well clear of both decision boundaries, most verdicts have room to survive outlier handling. RE = random-effects; UWLS = unrestricted weighted least squares.

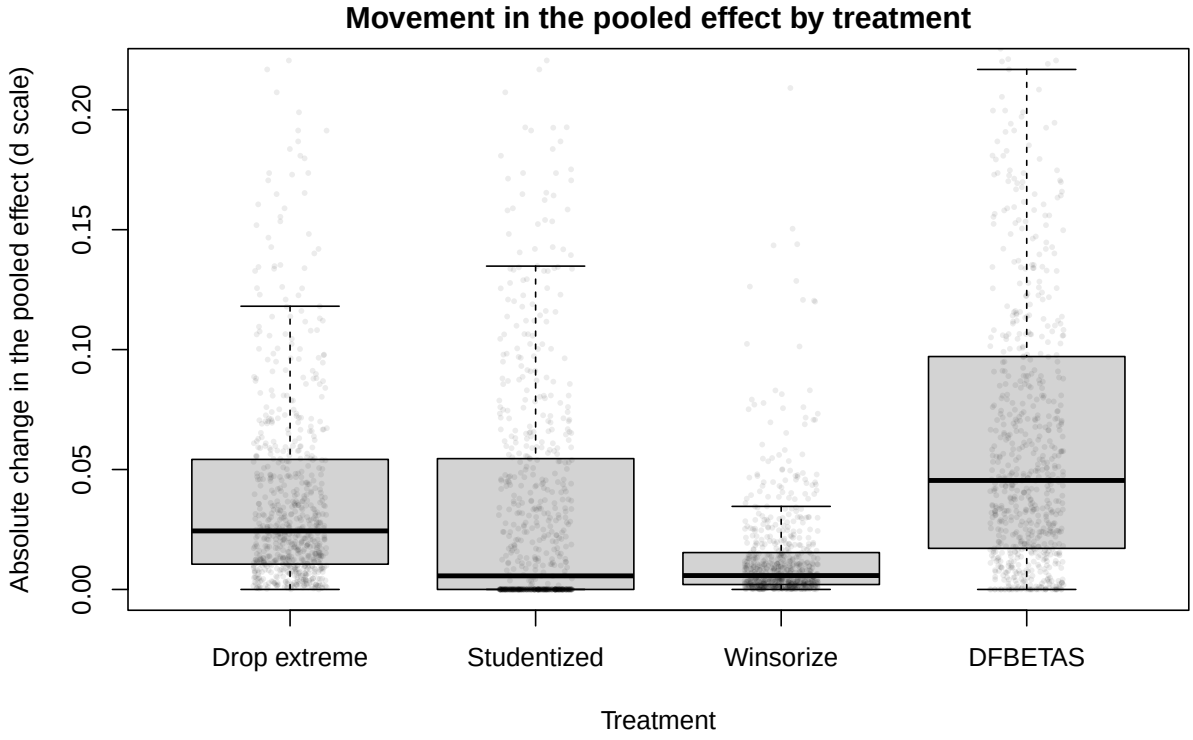

Figure A.1. Most outlier rules move the pooled effect very little; the movement is small and right-skewed for every rule. *Notes: Pre-registered figure. The unit is a meta-analysis-by-estimator result, pooled across both estimators (random-effects and unrestricted weighted least squares) and all three fields. Each box summarizes the absolute change in the pooled effect, $|\Delta d|$ (Cohen's d scale), that one of the four active rules produces relative to the untreated baseline: drop-extreme, studentized, winsorize, and DFBETAS. Boxes and whiskers show the interquartile range and adjacent values of $|\Delta d|$ (drawn without outlier points); the overlaid translucent dots are the individual meta-analysis-by-estimator cells, horizontally jittered and, for display only, clipped at the 99th percentile of $|\Delta d|$ so the extreme tail does not stretch the axis. The y-axis is the absolute change in the pooled effect on the d scale; the x-axis is the four rules.*

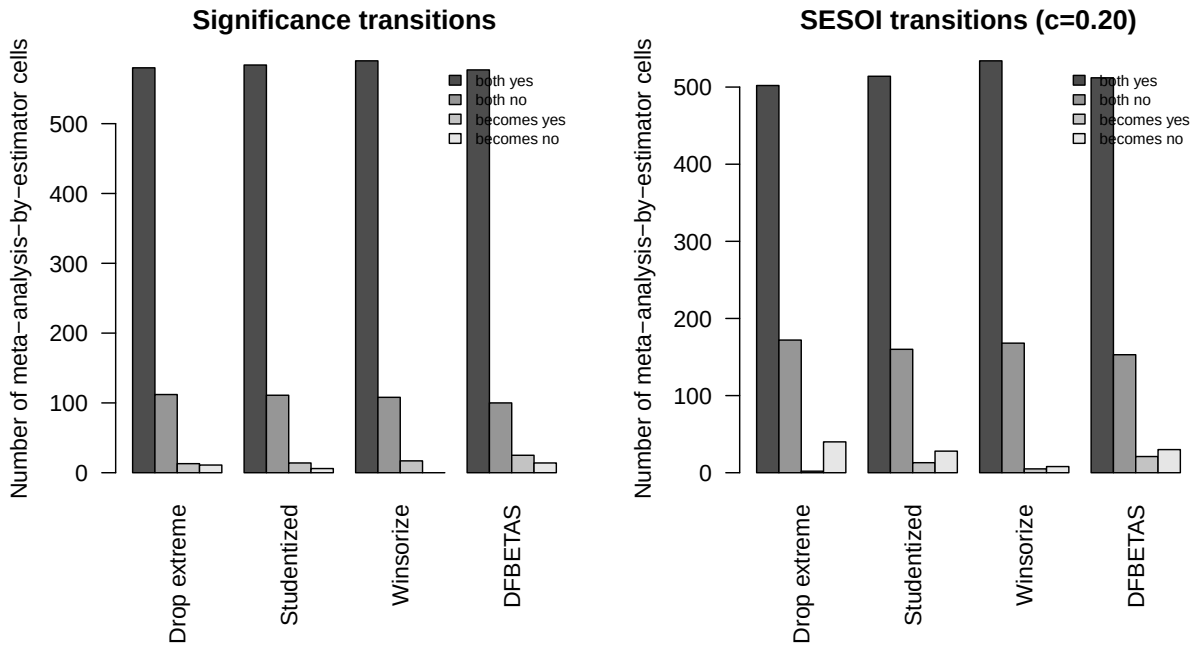

Figure A.2. Verdict transitions are dominated by agreement under every rule. *Notes: Pre-registered figure. Counts are meta-analysis-by-estimator cells, pooled over both estimators (random-effects and unrestricted weighted least squares) and all three fields, so each rule's bars sum to about* 716 *(minus any non-estimable cells). The left panel, "Significance transitions," classifies each cell by how its statistical-significance verdict (significant iff two-sided* $p < 0.05$*) changes from the untreated baseline to the rule; the right panel, "SESOI transitions (*$c = 0.20$*)," does the same for the smallest-effect-size-of-interest verdict (*$|d| \geq 0.20$*). Within each panel the bars are grouped by the four active rules (drop-extreme, studentized, winsorize, DFBETAS), and within each group the four bars are, in order, "both yes" (verdict positive before and after), "both no" (negative before and after), "becomes yes" (flips to positive), and "becomes no" (flips to negative). The y-axis counts meta-analysis-by-estimator cells.*

Table A.4. *Edge-case ledger: the handful of non-estimable cells, non-finite diagnostics, and boundary exclusions behind the main results.*

Edge case	Count
Not-estimable cells, main spec (of 3,580 table rows)	2
RE, studentized, Sladekova/A181_5 (psychology)	1
RE, winsorize, Sladekova/A181_5 (psychology)	1
Non-finite diagnostic rows (of 33,328), treated as unflagged	3
Retained large estimates, $\lvert d\rvert > 10$	37
Repeat rows, identical (d, se) value (rows / meta-analyses)	664 / 61

Note: A ledger of the boundary and edge cases that arise in the main specification. "Not-estimable cells" are meta-analysis-by-estimator-by-treatment cells with registered status other than estimable; both are the same meta-analysis (Sladekova/A181_5) under RE, failing with a model error at the studentized and winsorize treatments only (it is estimable under drop-extreme, DFBETAS, and no-treatment, and under UWLS throughout). There are no $k < 3$ non-estimable cells and none under UWLS. "Non-finite diagnostic rows" are per-estimate diagnostic rows with a non-finite externally studentized residual or DFBETAS (all three under RE); by the registered rule a non-finite diagnostic is treated as not flagged, so these estimates are never removed by the studentized or DFBETAS rules. "Retained large estimates, $\lvert d\rvert > 10$" are the 37 estimates with pooled-scale magnitude above 10 that pass the data-validity filter and are kept. The repeat rows carrying an identical (d, se) value (664 rows across 61 meta-analyses) are reported for reference. The Fisher-z boundary exclusion ($\lvert r\rvert \geq 0.999$) is applied at data-build time, before any committed output is written, so its count cannot be recovered from the frozen outputs; the retained $\lvert d\rvert > 10$ count is the nearest recoverable summary of that near-unit tail, not the excluded count itself. RE = random-effects; UWLS = unrestricted weighted least squares.

Table A.5. *Full four-cell agreement of each active rule with doing nothing, by outcome and estimator.*

Estimator	Rule	Both yes	Both no	Becomes yes	Becomes no	Not est.	n est.	Switch rate
Panel A. Significance ($p < 0.05$)								
RE	Drop-extreme	289	55	6	8	0	358	3.91%
	Studentized	292	55	6	4	1	357	2.80%
	Winsorize	296	52	9	0	1	357	2.52%
	DFBETAS	290	52	9	7	0	358	4.47%
UWLS	Drop-extreme	291	57	7	3	0	358	2.79%
	Studentized	292	56	8	2	0	358	2.79%
	Winsorize	294	56	8	0	0	358	2.23%
	DFBETAS	287	48	16	7	0	358	6.42%
Pooled	Drop-extreme	580	112	13	11	0	716	3.35%
	Studentized	584	111	14	6	1	715	2.80%
	Winsorize	590	108	17	0	1	715	2.38%
	DFBETAS	577	100	25	14	0	716	5.45%
Panel B. SESOI-0.20 ($\lvert d \rvert \geq 0.20$)								
RE	Drop-extreme	256	78	1	23	0	358	6.70%
	Studentized	266	76	2	13	1	357	4.20%
	Winsorize	274	77	1	5	1	357	1.68%
	DFBETAS	259	75	4	20	0	358	6.70%
UWLS	Drop-extreme	246	94	1	17	0	358	5.03%
	Studentized	248	84	11	15	0	358	7.26%
	Winsorize	260	91	4	3	0	358	1.96%
	DFBETAS	253	78	17	10	0	358	7.54%
Pooled	Drop-extreme	502	172	2	40	0	716	5.87%
	Studentized	514	160	13	28	1	715	5.73%
	Winsorize	534	168	5	8	1	715	1.82%
	DFBETAS	512	153	21	30	0	716	7.12%

Note: Each cell is a meta-analysis-by-estimator result; the four count columns cross the treated verdict against the untreated (do-nothing) verdict. Panel A: significant iff $p < 0.05$. Panel B: SESOI reached iff the registered rule $\lvert d \rvert \geq 0.20$ holds. "Both yes" = the verdict holds under both the rule and none; "Both no" = it holds under neither; "Becomes yes" = the rule newly turns the verdict on; "Becomes no" = the rule turns it off; the two off-diagonal "Becomes" columns are switches relative to doing nothing. "Not est." = cells not estimable under that rule; n est. is the estimable denominator; the switch rate is (Becomes yes + Becomes no) / n est. Pooled rows sum RE and UWLS. RE = random-effects; UWLS = unrestricted weighted least squares; drop-extreme, studentized, winsorize, and DFBETAS are the four active handling rules.

Table A.6. *DFBETAS under UWLS deletes estimates carrying by far the largest share of meta-analytic weight.*

Rule	Estimator	Cells altered	Median weight share	Median k share
Drop-extreme	RE	358	1.72%	4.8%
	UWLS	358	1.72%	4.8%
Studentized	RE	177	2.01%	4.2%
	UWLS	193	6.37%	4.3%
Winsorize	RE	356	7.75%	12.5%
	UWLS	357	7.75%	12.5%
DFBETAS	RE	335	5.71%	7.1%
	UWLS	346	23.17%	8.3%

Note: Exploratory. For each rule × estimator, the sample is the meta-analysis-by-estimator cells in which the rule alters at least one estimate; "Cells altered" counts them. "Median weight share" is the median across those cells of the share of inverse-variance weight ($w = 1/\text{se}^2$, within a meta-analysis) carried by the estimates the rule flags, deletes, or caps; "Median k share" is the median share of the meta-analysis's k estimates so affected. RE = random-effects; UWLS = unrestricted weighted least squares. Drop-extreme deletes the single registered most-extreme estimate; studentized flags |studentized residual| > 3; winsorize caps estimates beyond the 5th/95th within-meta-analysis quantiles; DFBETAS flags estimates whose deletion moves the pooled effect by more than the standard cutoff. Drop-extreme and winsorize identify the same estimates regardless of estimator (hence RE = UWLS). DFBETAS under UWLS removes a median of about 23% of the inverse-variance weight (interquartile range 12.7–37.3%) from a median of roughly 8% of a meta-analysis's estimates, the largest share of meta-analytic weight removed by any rule, consistent with UWLS switching more often than RE and with DFBETAS moving pooled results the most.

Table A.7. *UWLS switches more often than RE on both outcomes, and DFBETAS switches the largest share of verdicts of any single rule.*

	Significance ($p < 0.05$)	SESOI-0.20 ($\lvert d\rvert \geq 0.20$)
Panel A. ≥ 1 rule switches, by estimator (cell level, $n = 715$)		
RE-only	6.44%	8.96%
UWLS-only	8.94%	11.73%
Panel B. ≥ 1 rule switches, either estimator (meta-analysis level, $n = 358$)		
Either estimator	11.45%	15.92%
Panel C. Single rule vs. doing nothing (cell level, $n = 715$)		
Drop-extreme	3.36%	5.87%
Studentized	2.80%	5.73%
Winsorize	2.38%	1.82%
DFBETAS	5.31%	7.13%

Note: Switching rates relative to doing nothing. A result is significant when $p < 0.05$; the SESOI-0.20 verdict uses the registered rule $|d| \geq 0.20$. Panels A and C use the meta-analysis-by-estimator cell as the unit, restricted to the 715 cells estimable under all five treatments; Panel B uses the meta-analysis as the unit ($n = 358$), counting a meta-analysis as switching if its RE or its UWLS cell switches. UWLS switches more often than RE on both outcomes (Panel A). Among single rules (Panel C), DFBETAS switches the largest share of verdicts and winsorize the fewest. Inclusion sensitivity: adding the one otherwise-excluded cell (Sladekova/A181_5 under RE) moves the pooled significance switching rate from 7.69% to 7.82%. RE = random-effects; UWLS = unrestricted weighted least squares.

Table A.8. *In one illustrative meta-analysis, only DFBETAS switches the significance verdict.*

Treatment	k after	Estimates changed	Pooled $\lvert d\rvert$	Two-sided p
None	18	0	0.318	0.017
Drop-extreme	17	1	0.258	0.032
Studentized	18	0	0.318	0.017
Winsorize	18	2	0.313	0.015
DFBETAS	16	2	0.179	0.082

Note: A single worked meta-analysis (Sladekova/B356_1; psychology; $k = 18$; $I^2 = 70.2\%$), refit under random effects (RE) with the Hartung–Knapp adjustment. The pooled effect is negative throughout; the table reports its absolute value $\lvert d\rvert$. "k after" is the number of estimates remaining after the treatment (drop-extreme removes one; DFBETAS removes two); "Estimates changed" counts the estimates deleted or capped. A result is significant when $p < 0.05$. Only DFBETAS crosses that threshold here, moving p from 0.017 to 0.082 as it deletes two estimates and pulls $\lvert d\rvert$ from 0.318 to 0.179. This case illustrates how a switch can arise. It is not used to estimate how often switches occur.

Table A.9. *The SESOI switching rate peaks near $c = 0.25$–0.30, where the pooled-effect mass sits, and reproduces the headline at $c = 0.20$.*

c	≥ 1 rule crosses	All five identical	c	≥ 1 rule crosses	All five identical
0.05	7.83%	92.17%	0.30	12.31%	87.69%
0.10	6.43%	93.57%	0.35	9.65%	90.35%
0.15	7.13%	92.87%	0.40	8.53%	91.47%
0.20	**10.35%**	**89.65%**	0.45	6.15%	93.85%
0.25	12.45%	87.55%	0.50	6.43%	93.57%

Note: The $c = 0.20$ (main) and $c = 0.10$ (sensitivity) rows are the pre-registered SESOI thresholds; the remaining rows extend the grid as exploratory context. Unit is the meta-analysis-by-estimator cell estimable under all five treatments ($n = 715$, over both RE and UWLS). A smallest-effect-size-of-interest (SESOI) verdict is reached when the registered rule $\lvert d\rvert \geq c$ holds. "≥ 1 rule crosses" is the share of cells where at least one of the four active rules (drop-extreme, studentized, winsorize, DFBETAS) flips the $\lvert d\rvert \geq c$ verdict relative to doing nothing; "All five identical" is the complementary share where none plus all four active rules agree. The main-text SESOI threshold $c = 0.20$ (10.35%) reproduces the headline crossing rate; switching peaks near $c = 0.25$–0.30 and is lowest at the extremes. RE = random-effects; UWLS = unrestricted weighted least squares.