Reproduce Zhao's results

Re-extracted

Preemptive response

Verification of Zhao et al. (2023)

Code

Ian Hussey

Show

Reproduce Zhao's results

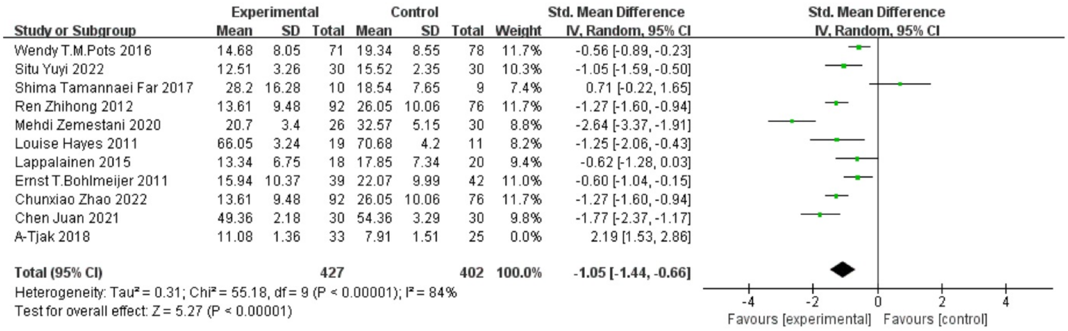
Zhao et al. (2023) present the results of multiple meta-analyses, sensitivity analyses, and subgroup analyses. Here, I attempt to reproduce and verify the result of only the first meta-analysis reported in the first forest plot (figure 4), which included the largest number of studies and data for the original studies' post-intervention time point.

Zhao et al. (2023, p. 6) reported "Effect of ACT on depression: Eleven studies reported the effect of ACT on depression levels in 887 patients with depressive disorders. There was statistically significant heterogeneity among the studies (P < 0.00001; I2= 93%), therefore, the random effects model was used to conduct meta-analysis. There was significant treatment effect from ACT in comparison to other treatments, as shown in the forest plot in Fig. 4."

However, Figure 4 does not present the full meta-analysis, but rather a sensitivity analysis after excluding A-Tjak et al. (2018). The rationale for excluding that study is not stated. I therefore first fit a meta-analysis to all 11 studies, then perform a series of outlier/undue influence tests to try to reproduce the conclusion that A-Tjak et al. (2018) should be excluded, and then finally I fit a new meta-analysis to the remaining 10 studies to try to reproduce the meta-analytic effect reported in Zhao et al. (2023) Figure 4.

Summary stats (M, SD, N for both groups) were extracted from Zhao et al.'s (2023) forest plot in Figure 4. Specifically:

Show



	Experimental			Control				Std. Mean Difference
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Random, 95% CI
Wendy T.M.Pots 2016	14.68	8.05	71	19.34	8.55	78	11.7%	-0.56 [-0.89, -0.23]
Situ Yuyi 2022	12.51	3.26	30	15.52	2.35	30	10.3%	-1.05 [-1.59, -0.50]
Shima Tamannaei Far 2017	28.2	16.28	10	18.54	7.65	9	7.4%	0.71 [-0.22, 1.65]
Ren Zhihong 2012	13.61	9.48	92	26.05	10.06	76	11.7%	-1.27 [-1.60, -0.94]
Mehdi Zemestani 2020	20.7	3.4	26	32.57	5.15	30	8.8%	-2.64 [-3.37, -1.91]
Louise Hayes 2011	66.05	3.24	19	70.68	4.2	11	8.2%	-1.25 [-2.06, -0.43]
Lappalainen 2015	13.34	6.75	18	17.85	7.34	20	9.4%	-0.62 [-1.28, 0.03]
Ernst T.Bohlmeijer 2011	15.94	10.37	39	22.07	9.99	42	11.0%	-0.60 [-1.04, -0.15]
Chunxiao Zhao 2022	13.61	9.48	92	26.05	10.06	76	11.7%	-1.27 [-1.60, -0.94]
Chen Juan 2021	49.36	2.18	30	54.36	3.29	30	9.8%	-1.77 [-2.37, -1.17]
A-Tjak 2018	11.08	1.36	33	7.91	1.51	25	0.0%	2.19 [1.53, 2.86]
Total (95% CI)			**427**			**402**	**100.0%**	**-1.05 [-1.44, -0.66]**

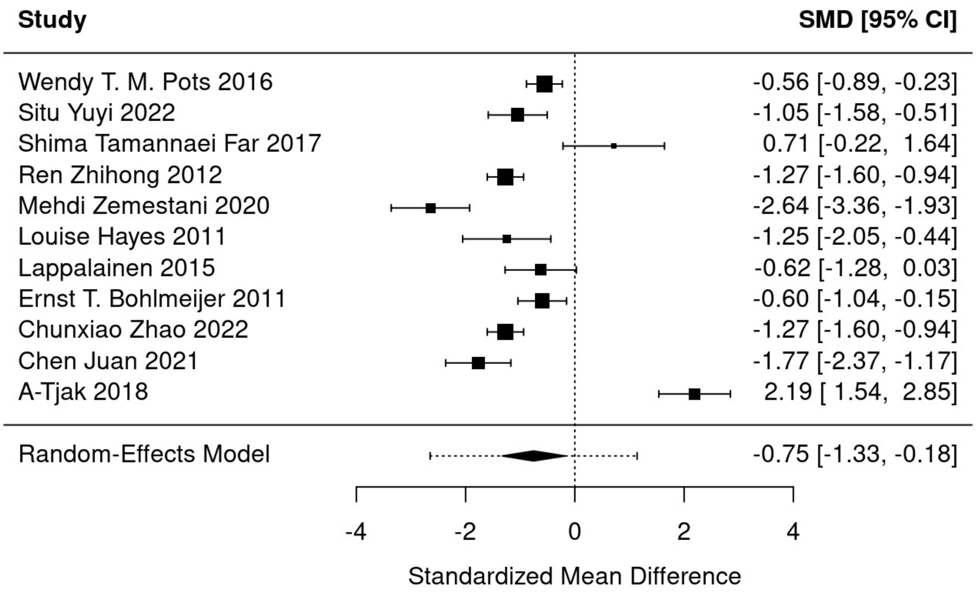
Heterogeneity: Tau² = 0.31; Chi² = 55.18, df = 9 (P < 0.00001); I² = 84%

Test for overall effect: Z = 5.27 (P < 0.00001)

Fig. 4 Effect of ACT on depression in patients with depressive disorders after sensitivity analysis

Meta-analysis of all 11 studies

Show



Show

```
##
## Random-Effects Model (k = 11; tau^2 estimator: DL)
##
## tau^2 (estimated amount of total heterogeneity): 0.8485 (SE = 0.4826)
## tau (square root of estimated tau^2 value):      0.9212
## I^2 (total heterogeneity / total variability):   93.11%
## H^2 (total variability / sampling variability):  14.51
##
## Test for Heterogeneity:
## Q(df = 10) = 145.0800, p-val < .0001
##
## Model Results:
##
## estimate      se     zval    pval    ci.lb    ci.ub
##  -0.7548  0.2925  -2.5809  0.0099  -1.3280  -0.1816  **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heterogeneity: Q(df = 10) = 145.08, p < .0001, tau^2 = 0.8485, I^2 = 93.1%, H^2 = 14.51

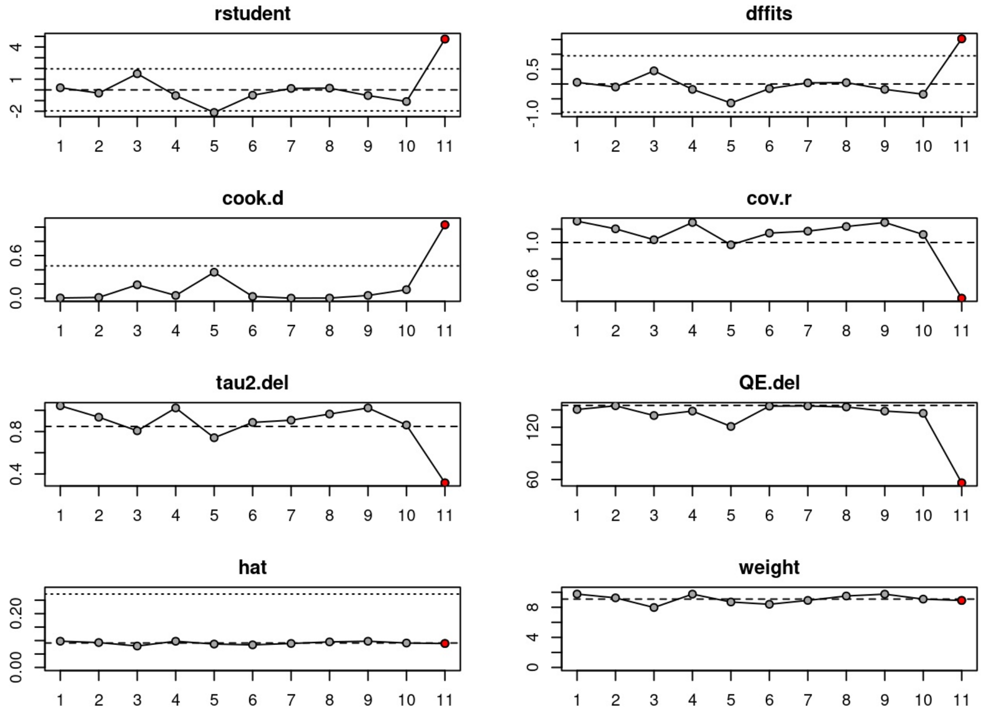
Outlier/undue influence tests

No specific justification was given in Zhao et al. (2023) for their exclusion of A-Tjak et al. (2018). In the absence of this, I employ {metafor}'s `influence()` function which applies multiple different metrics of undue influence.

Show

```
##
##                            rstudent dffits cook.d cov.r tau2.del QE.del  hat
## Wendy T. M. Pots 2016        0.20   0.06   0.00  1.34     1.04 140.42 0.10
## Situ Yuyi 2022              -0.30  -0.10   0.01  1.20     0.94 144.70 0.09
## Shima Tamannaei Far 2017     1.51   0.44   0.19  1.04     0.81 133.49 0.08
## Ren Zhihong 2012            -0.53  -0.18   0.04  1.31     1.02 138.61 0.10
## Mehdi Zemestani 2020        -2.11  -0.64   0.37  0.97     0.74 120.95 0.09
## Louise Hayes 2011           -0.50  -0.15   0.02  1.14     0.89 144.26 0.08
## Lappalainen 2015             0.13   0.04   0.00  1.17     0.91 144.45 0.09
## Ernst T. Bohlmeijer 2011     0.16   0.05   0.00  1.24     0.97 143.33 0.10
## Chunxiao Zhao 2022          -0.53  -0.18   0.04  1.31     1.02 138.61 0.10
## Chen Juan 2021              -1.09  -0.34   0.12  1.12     0.86 136.07 0.09
## A-Tjak 2018                  4.74   1.52   1.03  0.47     0.32  56.20 0.09
##                            weight  dfbs inf
## Wendy T. M. Pots 2016        9.76  0.06
## Situ Yuyi 2022               9.25 -0.10
## Shima Tamannaei Far 2017     7.97  0.44
## Ren Zhihong 2012             9.75 -0.18
## Mehdi Zemestani 2020         8.70 -0.64
## Louise Hayes 2011            8.40 -0.15
## Lappalainen 2015             8.92  0.04
## Ernst T. Bohlmeijer 2011     9.50  0.05
## Chunxiao Zhao 2022           9.75 -0.18
## Chen Juan 2021               9.09 -0.34
## A-Tjak 2018                  8.91  1.55    *
```

Show

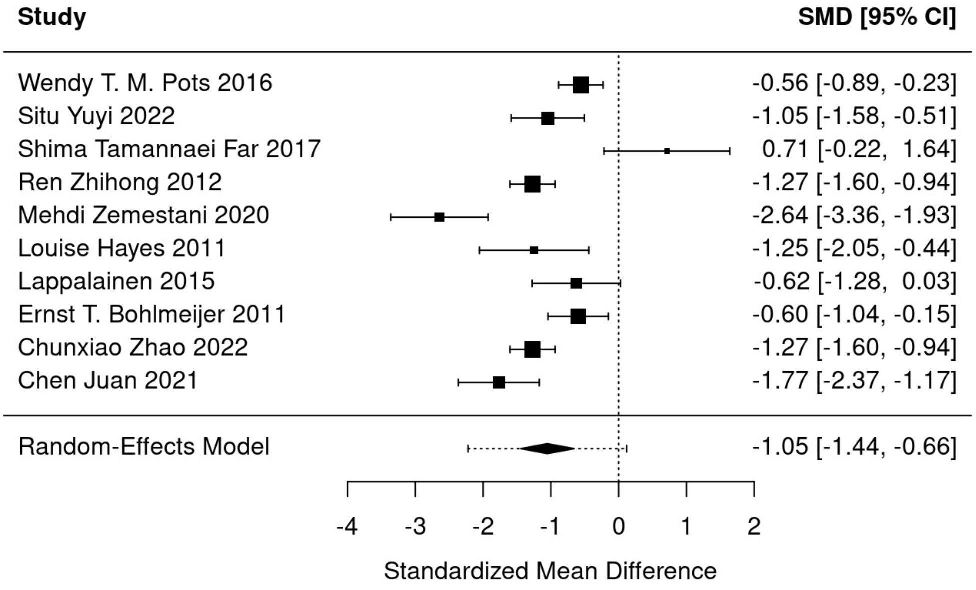


A-Tjak was detected as an outlier. Although Zhao et al.'s (2023) specific method of outlier detection is unknown, their conclusion is replicate by these results.

Sensitivity analysis: Exclude Tjak et al. (2018)

As in Zhao Figure 4, and following the results of the above outlier tests, I exclude A-Tjak et al. (2018) and fit a new meta-analysis in order to attempt to reproduce the results reported in Figure 4.

Show



Show

```
##
## Random-Effects Model (k = 10; tau^2 estimator: DL)
##
## tau^2 (estimated amount of total heterogeneity): 0.3159 (SE = 0.2036)
## tau (square root of estimated tau^2 value):      0.5621
## I^2 (total heterogeneity / total variability):   83.99%
## H^2 (total variability / sampling variability):  6.24
##
## Test for Heterogeneity:
## Q(df = 9) = 56.2016, p-val < .0001
##
## Model Results:
##
## estimate      se     zval    pval    ci.lb    ci.ub
##  -1.0520  0.2004  -5.2504  <.0001  -1.4447  -0.6593  ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zhao et al.'s (2023) forest plot in Figure 4 reported: Meta-analytic d = -1.05, 95% CI [-1.44, -0.66]. The results above therefore reproduce this result from the summary statistics reported in that forest plot.

Heterogeneity: Q(df = 9) = 56.2, p < .0001, tau^2 = 0.3159, I^2 = 84%, H^2 = 6.24

NNT

NNT = 1.84, 95% CI [1.44, 2.79]

Re-extracted

Although the results reported in Figure 4 can be reproduced from the summary statistics reported in Figure 4, the validity of those results are reliant on the summary statistics being accurately extracted from the original studies. Recent work has shown that this is frequently not the case (e.g., Maassen et al., 2020). I therefore attempted to extract the Mean, SD, and N for the ACT and control conditions at the post-intervention time point for each of the original studies.

In the below plots, I use the same labelling as in Zhao et al. (2023). The references (including DOIs) for the studies they refer to are available in the excel files that the data is loaded from. Note that there is imperfect correspondence between Zhao's labels and the authors of the original studies.

Comparisons of extracted summary stats

Show

comparison	n
Consistent	5
Duplicate	1
Inconsistent	3
Indeterminable	2

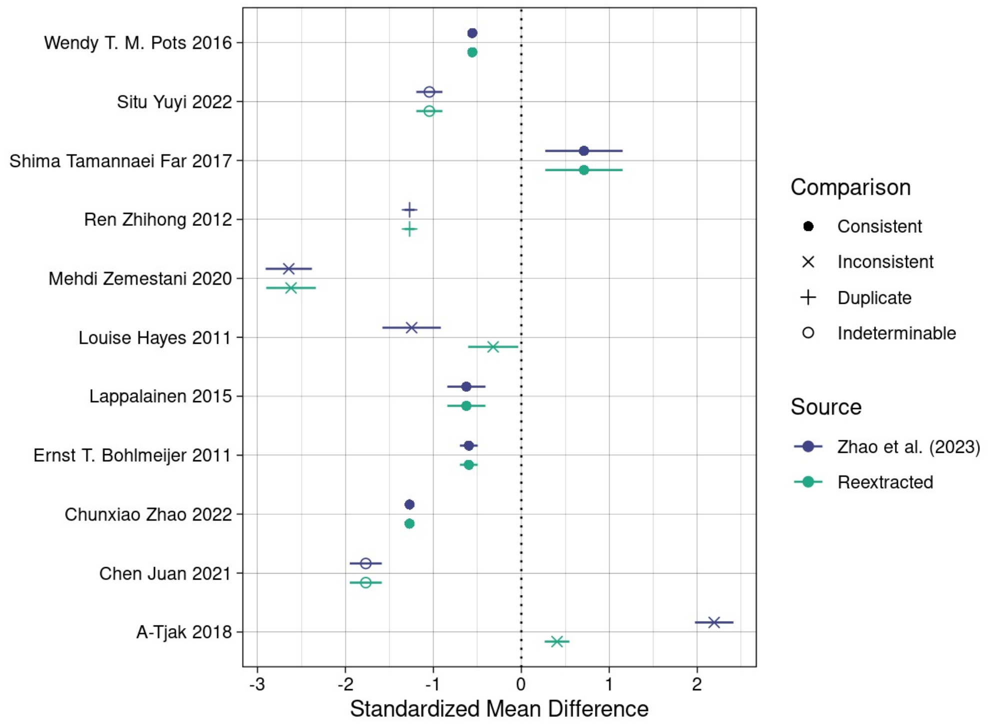
Two articles with DOIs were only available in Chinese for a fee, and were therefore not accessed. One article did not have a DOI listed and could not be found via internet searches. It is possible that the reference listed in Zhao et al. (2023) is an English translation of the original Chinese title. As such, the reproducibility of 3 of the 11 studies was indeterminable.

In 3 of the remaining 8 articles, the summary statistics (M, SD, N) reported in Zhao et al. (2023, Figure 4) do not match those reported in the original studies. In two cases (Louise Hayes 2011, A-Tjak 2018), the original articles reported Standard Errors and these were employed by Zhao et al. (2023) as if they were Standard Deviations. This is a common and unfortunate issue in meta analyses, as is problematic as SEs are much smaller than SDs, therefore inflating the apparent Standardized Mean Difference effect sizes employed in the meta-anaalysis. In one case (Mehdi Zemestani 2020), the Ns do not match those reported in the original publication, although the nature of the error is unclear. As such, only 5 of the 11 sets of summary statistics could be verified against the numbers reported in the original articles.

Show

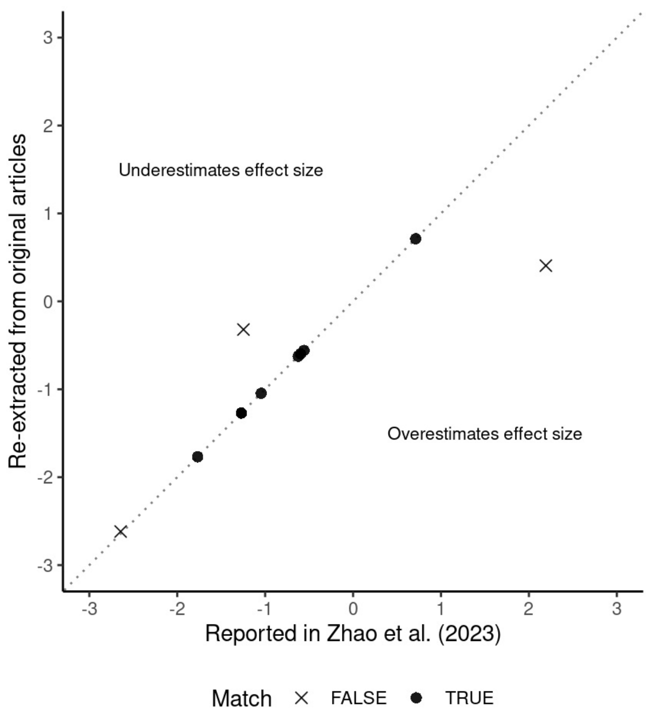
study	source	mean_exp	sd_exp	n_exp	mean_control	sd_control	n_control	reextracted	comparison
A-Tjak 2018	Zhao et al. (2023)	11.08	1.36	33	7.91	1.51	25	NA	
A-Tjak 2018	Reextracted	11.08	7.81	33	7.91	7.55	25	TRUE	Inconsistent
Chen Juan 2021	Zhao et al. (2023)	49.36	2.18	30	54.36	3.29	30	NA	
Chen Juan 2021	Reextracted	49.36	2.18	30	54.36	3.29	30	FALSE	Indeterminable
Chunxiao Zhao 2022	Zhao et al. (2023)	13.61	9.48	92	26.05	10.06	76	NA	
Chunxiao Zhao 2022	Reextracted	13.61	9.48	92	26.05	10.06	76	TRUE	Consistent
Ernst T. Bohlmeijer 2011	Zhao et al. (2023)	15.94	10.37	39	22.07	9.99	42	NA	
Ernst T. Bohlmeijer 2011	Reextracted	15.94	10.37	39	22.07	9.99	42	TRUE	Consistent
Lappalainen 2015	Zhao et al. (2023)	13.34	6.75	18	17.85	7.34	20	NA	
Lappalainen 2015	Reextracted	13.34	6.75	18	17.85	7.34	20	TRUE	Consistent
Louise Hayes 2011	Zhao et al. (2023)	66.05	3.24	19	70.68	4.20	11	NA	
Louise Hayes 2011	Reextracted	66.05	14.12	19	70.68	13.93	11	TRUE	Inconsistent
Mehdi Zemestani 2020	Zhao et al. (2023)	20.70	3.40	26	32.57	5.15	30	NA	
Mehdi Zemestani 2020	Reextracted	20.70	3.40	23	32.57	5.15	29	TRUE	Inconsistent
Ren Zhihong 2012	Zhao et al. (2023)	13.61	9.48	92	26.05	10.06	76	NA	
Ren Zhihong 2012	Reextracted	13.61	9.48	92	26.05	10.06	76	FALSE	Duplicate
Shima Tamannaei Far 2017	Zhao et al. (2023)	28.20	16.28	10	18.54	7.65	9	NA	
Shima Tamannaei Far 2017	Reextracted	28.20	16.28	10	18.54	7.65	9	TRUE	Consistent
Situ Yuyi 2022	Zhao et al. (2023)	12.51	3.26	30	15.52	2.35	30	NA	
Situ Yuyi 2022	Reextracted	12.51	3.26	30	15.52	2.35	30	FALSE	Indeterminable
Wendy T. M. Pots 2016	Zhao et al. (2023)	14.68	8.05	71	19.34	8.55	78	NA	
Wendy T. M.	Reextracted	14.68	8.05	71	19.34	8.55	78	TRUE	Consistent

Show



Show

Show

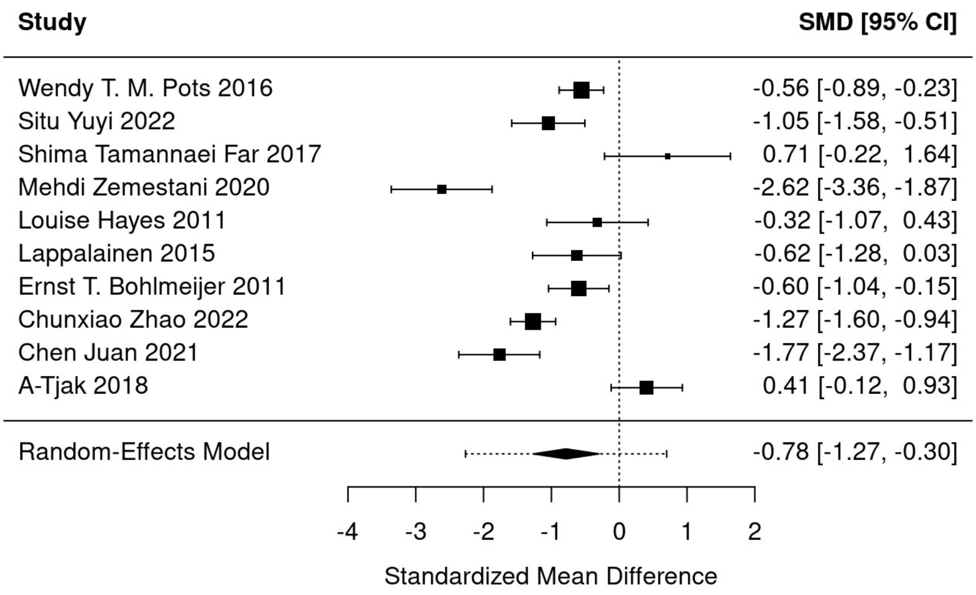


Show

Updated meta-analysis of the 10 studies after excluding the duplicate

Following the same workflow as Zhao et al. (2023), a random effects meta-analysis was then fit to the 10 sets of effect sizes (i.e., including the 3 that were indeterminable, on the liberal assumption that they are correct, but excluding the duplicate on the reasonable assumption that it is not).

Show



Show

```
##
## Random-Effects Model (k = 10; tau^2 estimator: DL)
##
## tau^2 (estimated amount of total heterogeneity): 0.5125 (SE = 0.3143)
## tau (square root of estimated tau^2 value):      0.7159
## I^2 (total heterogeneity / total variability):   88.28%
## H^2 (total variability / sampling variability):  8.53
##
## Test for Heterogeneity:
## Q(df = 9) = 76.7813, p-val < .0001
##
## Model Results:
##
## estimate      se     zval    pval    ci.lb    ci.ub
##  -0.7831  0.2459  -3.1848  0.0014  -1.2651  -0.3012  **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Heterogeneity: Q(df = 9) = 76.78, p < .0001, tau^2 = 0.5125, I^2 = 88.3%, H^2 = 8.53

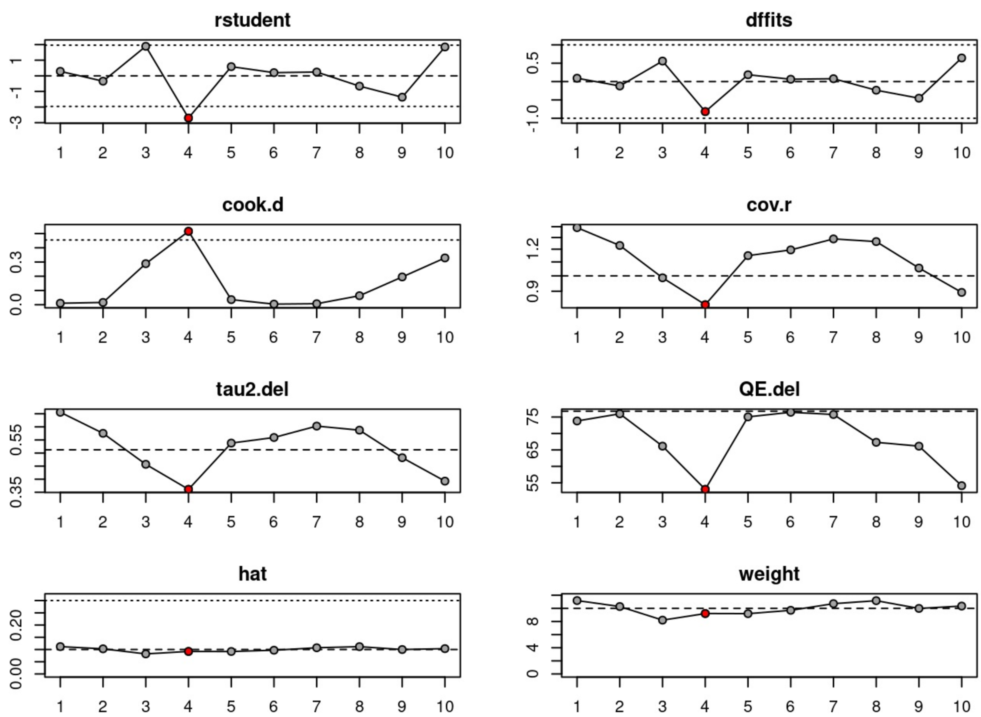
Outlier/undue influence tests

Similarly, I then checked for outliers/undue influence.

Show

```
##
##                           rstudent dffits cook.d cov.r tau2.del QE.del  hat
## Wendy T. M. Pots 2016         0.29   0.09   0.01  1.39     0.66  73.81 0.11
## Situ Yuyi 2022               -0.34  -0.12   0.02  1.23     0.58  75.98 0.10
## Shima Tamannaei Far 2017      1.89   0.56   0.29  0.99     0.46  66.15 0.08
## Mehdi Zemestani 2020         -2.70  -0.82   0.52  0.82     0.36  53.00 0.09
## Louise Hayes 2011             0.59   0.19   0.04  1.15     0.54  75.05 0.09
## Lappalainen 2015              0.20   0.06   0.00  1.19     0.56  76.45 0.10
## Ernst T. Bohlmeijer 2011      0.24   0.08   0.01  1.29     0.60  75.77 0.11
## Chunxiao Zhao 2022           -0.66  -0.24   0.06  1.26     0.59  67.31 0.11
## Chen Juan 2021               -1.37  -0.45   0.20  1.05     0.48  66.14 0.10
## A-Tjak 2018                   1.85   0.64   0.33  0.89     0.39  54.12 0.10
##                           weight  dfbs inf
## Wendy T. M. Pots 2016     11.19  0.09
## Situ Yuyi 2022            10.28 -0.12
## Shima Tamannaei Far 2017   8.20  0.56
## Mehdi Zemestani 2020       9.21 -0.83   *
## Louise Hayes 2011          9.19  0.19
## Lappalainen 2015           9.70  0.06
## Ernst T. Bohlmeijer 2011  10.72  0.08
## Chunxiao Zhao 2022        11.17 -0.24
## Chen Juan 2021             9.99 -0.45
## A-Tjak 2018               10.35  0.64
```

Show



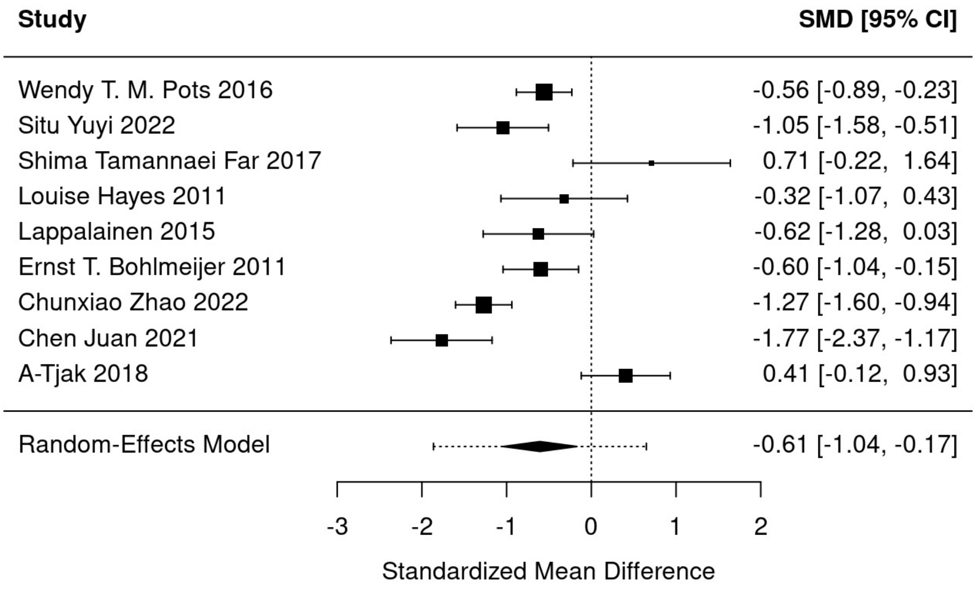
This suggests that Zemestani et al. (2020) is an outlier that should be excluded, whereas Zhao et al. (2023) excluded A-Tjak et al. (2018).

Note that there are additional reasons to exclude the results of Zemestani et al. (2020): as documented in this pubpeer comment (https://pubpeer.com/publications/0E13E34679B18385D6C4C29143A9CD), several results reported in that article fail StatCheck, GRIM, and GRIMMER tests (i.e., the reported results are mathmatically impossible); the reported SD of the BDI-II scores is implausibly small given normative data for the BDI-II elsewhere, and as a result the SMD effect size is implausibly large (d = -2.64 in Zhao et al. 2023).

Sensitivity analysis: Exclude Zemestani et al. (2020)

I therefore fitted a new meta-analysis to the remaining 9 studies after excluding Zemestani et al. (2020).

Show



Show

```
##
## Random-Effects Model (k = 9; tau^2 estimator: DL)
##
## tau^2 (estimated amount of total heterogeneity): 0.3613 (SE = 0.2398)
## tau (square root of estimated tau^2 value):      0.6011
## I^2 (total heterogeneity / total variability):   84.90%
## H^2 (total variability / sampling variability):  6.62
##
## Test for Heterogeneity:
## Q(df = 8) = 52.9976, p-val < .0001
##
## Model Results:
##
## estimate      se     zval    pval    ci.lb    ci.ub
##  -0.6065  0.2227  -2.7230  0.0065  -1.0431  -0.1700  **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Show

A significant meta-analytic effect was still found, but this effect was -42.3% smaller than that reported by Zhao et al. (2023) after correcting the data extraction errors and refitting the meta-analysis and sensitivity analysis workflow.

Heterogeneity: Q(df = 8) = 53, p < .0001, tau^2 = 0.3613, I^2 = 84.9%, H^2 = 6.62

NNT

NNT = 3.01, 95% CI [1.85, 10.45]

Preemptive response

To preempt one possible response to these concerns: readers might argue something along the lines of "the conclusions of Zhao et al. (2023) are not substantially affected by the correction of these errors, and therefore there is no strong need for concern." However, any such reactions should be seen as worrisome rather than encouraging: meta-analyses are often argued to be at the top of the Evidence Pyramid, and their goal is to precisely estimate the true effect size beyond merely estimating whether the true effect size is non-zero or not. To retreat to the weaker position that the obtained effect size does not substantively matter is to undermine the point of doing the meta-analysis in the first place. Either the empirical estimates matter or they don't.

Finally, as discussed at the start of this document, no attempt was made to reproduce or attempt to understand the impact of the detected extraction errors on the other analyses reported in Zhao et al. (2023). It is likely that the subgroup analyses reported after Figure 4 will also be affected by the issues detected above. It is also possible that other issues exist, given that errors were made at the relatively simple data extraction phase. The results and conclusions of Zhao et al. (2023) should therefore be regarded with great skepticism given the issues already detected, and require careful revision throughout to ensure no other errors were made.

Equally, this verification does not attempt to critique things that were not done by Zhao et al. (2023). For example, the authors do not apply any method of bias correction for p-hacking or publication bias, and instead take the reported results at face value. Applying such tests typically shrinks the observed estimate. Equally, although the authors did include a risk of bias assessment, the results of that assessment were not employed in any sensitivity analysis (e.g., using bias scores as weights or as a method of exclusion).