# Trustworthiness assessment of Gloster et al. (2020) 'Treating treatment non-responders: A meta-analysis of randomized controlled psychotherapy trials'

Ian Hussey

**Abstract**

Gloster et al. (2020) presents the first meta-analysis of the efficacy of psychotherapy in treatment-resistant clients. Here, I present an assessment of the trustworthiness of the results presented in Gloster et al. (2020), following previous work on assessing the trustworthiness of original research (e.g., Wilkinson et al., 2024, Wilkinson et al., 2025) and meta-analyses (e.g., Hussey, 2025; Maassen et al., 2020). Serious concerns are raised about the plausibility of the magnitude of the effect sizes included in the meta-analyses and the correct extraction of effect sizes from original studies. The findings arguably presented below represent clear evidence of major errors that compromise the reliability of the research findings presented in Gloster et al. (2020), therefore, following COPE guidelines, the work requires substantial correction or possibly retraction (COPE, 2019).

## Issue 1: implausibly large effect sizes

SMD effect sizes (i.e., Cohen's *d*) of between 5 and 8 are extremely implausible in any RCT for psychotherapy. They are especially so here given that 1) these RCTs focus on treatment resistant patients and b) the SMD of 8 is for a one year follow up, so the improvements are not merely massive but also sustained long term.

The range of the scale and estimates of its SD also suggest an SMD of 8 is extremely implausible. In clinical samples, the YMRS has an SD of about 4.5 (N = 211 patients, Targum et al. 2018; N = 209, Suppes et al. 2016). SMD = 8 means 8 SDs between the conditions, i.e., 4.5*8 = c.36 scale points difference. The YMRS scale is a 0-60 scale, 36 points is a difference of 60% of the scale's range on average, in a treatment resistant population, and at 1 year follow up.

Taken at face value, this is either the most effective psychotherapy intervention in the history of psychology, or something is amiss.

The SMD information is presented in Gloster et al.'s (2020) forest plots, but it becomes more apparent if the plots do not censor the effect sizes, as they do in Gloster et al. (2020). Below I (successfully) computationally reproduce Gloster et al.'s (2020) RE meta-analysis results from their effect sizes and standard errors in order to re-plot them without censoring.

### Posttreatment time point

Gloster et al. (2020) RE meta analysis at post treatment time point: SMD = 0.818

Without censoring effect sizes in the forest plot:

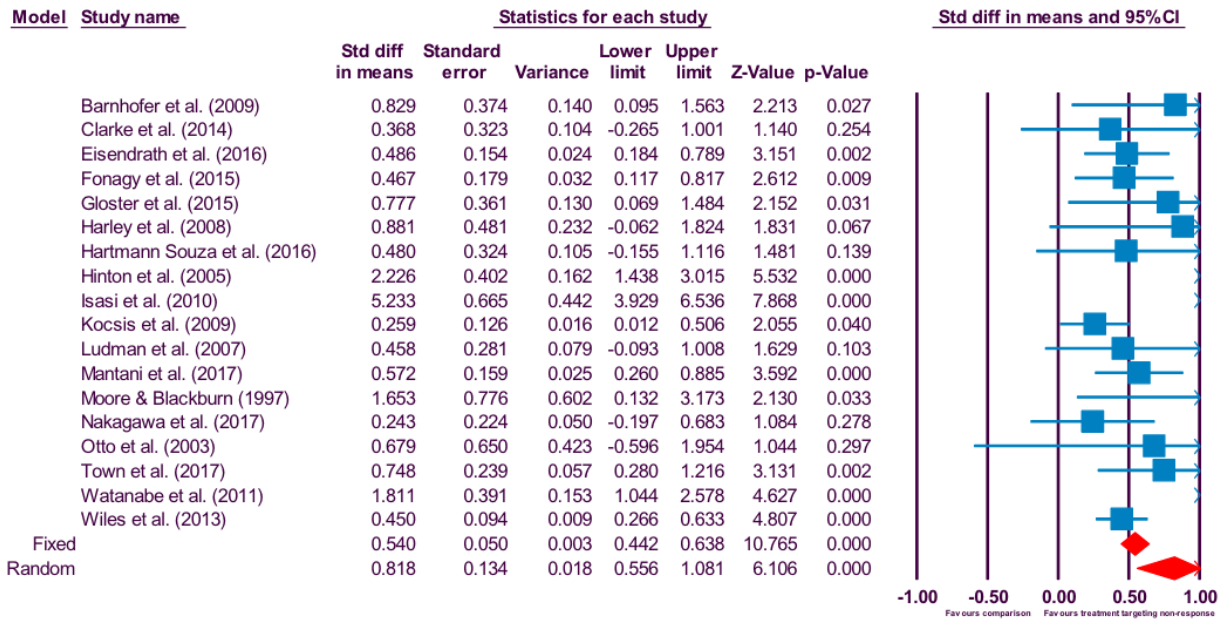
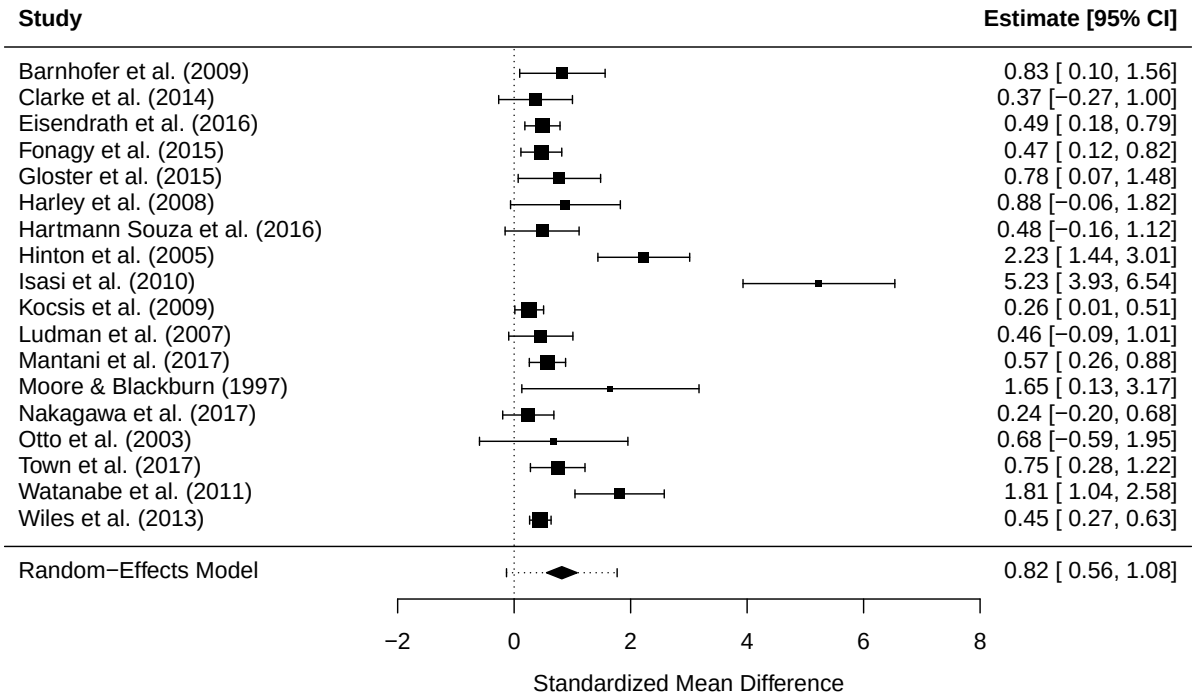



Figure 1: Gloster et al. (2020) Figure 2




- Reproduced successfully.
- Effect size for Isasi et al. (2010) is very implausibly large.
- Hinton et al. (2005), Moore & Blackburn (1997) and Watanabe et al. (2011) may also raise questions

about plausibility.

### Follow-up time point

Gloster et al. (2020) RE meta analysis at follow-up time point: SMD = 1.189

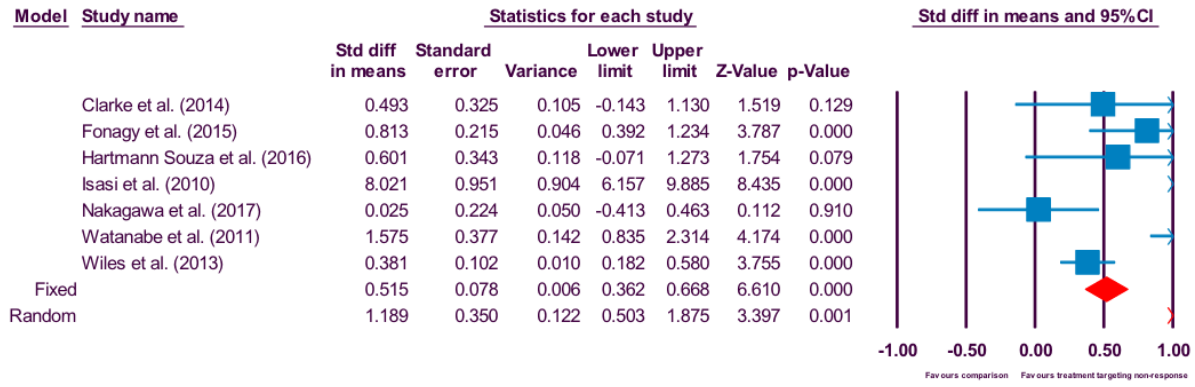
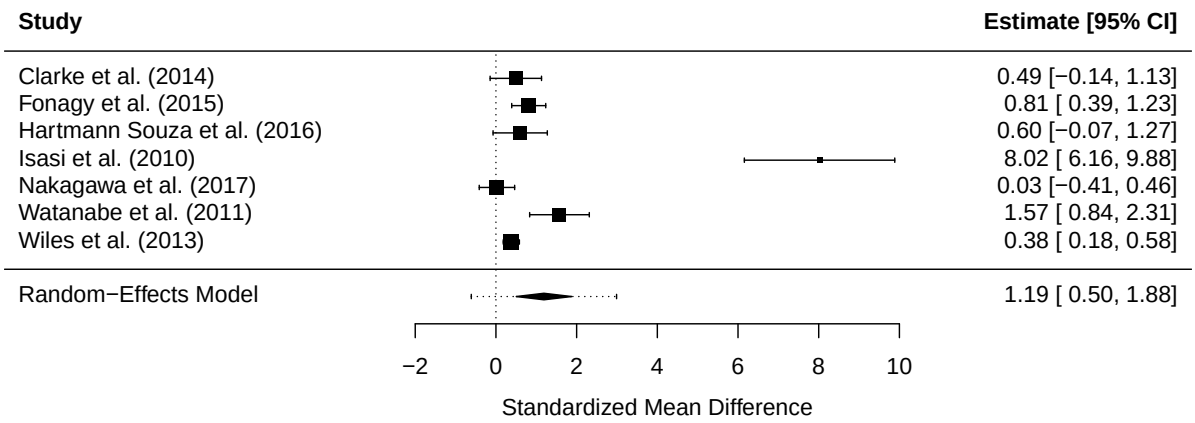



Figure 2: Gloster et al. (2020) Figure 2

Without censoring effect sizes in the forest plot:




- Reproduced successfully.
- Effect size for Isasi et al. (2010) is very implausibly large.
- Watanabe et al. (2011) may also raise questions about plausibility.

## Issue 2: effect size extraction errors

Implausibly large effect sizes can be caused by the confusion of Standard Errors and Standard Deviations, the former are much smaller than the latter (SD = SE * sqrt(N)). Given that the SMD effect size puts SDs in the denominator, misuse of SEs greatly increases the estimate. This error, the "Standard Error error", is known to be prevalent in published meta-analyses (eg Maassen et al., 2020).

Gloster et al. (2020) do not report the means and SDs they used to calculate the SMD effect sizes, only the Ns and the SMDs themselves. However, we can recompute them and compare the SMDs. I attempted to do this for the largest effect sizes in reported in Gloster et al.'s (2020) forest plots (figures 2 and 5).

## Isasi et al. (2010)

Gloster et al.'s (2020) forest plots (figures 2 and 5) include SMDs for Isasi et al. (2010) of 5.233 for postintervention and 8.02 for follow up.

Inspection of Isasi et al. (2010) shows that they reported means and "te"s, whose usage is atypical and meaning is unclear:

**Table 2**
Between-group differences at the different evaluation time points.

| Values scales | Pretreatment Mean (te) | (*p* value) [a] | Postreatment Mean (te) | (*p* value) [a] | 6 months Mean (te) | (*p* value) [a] | 12 months Mean (te) | (*p* value) [a] |
|---|---|---|---|---|---|---|---|---|
| *Number recent hospitalizations* | | | | | | | | |
| Control group | 0.39 (0.15) | | 0.28 (0.09) | | 0.33 (0.12) | | 0.39 (0.10) | |
| Experimental group | 0.10 (0.14) | (0.164) | 0.00 (0.09) | (0.037) * | 0.05 (0.11) | (0.089) | 0.00 (0.09) | (0.007) * |
| *Anxiety (STAI-S)* | | | | | | | | |
| Control group | 17.50 (2.43) | | 23.05 (2.82) | | 26.05 (3.00) | | 32.78 (2.67) | |
| Experimental group | 21.30 (2.30) | (0.264) | 16.00 (2.67) | (0.077) | 16.50 (2.85) | (0.027) * | 8.85 (2.53) | (<0.001) ** |
| *Beck depression (BDI)* | | | | | | | | |
| Control group | 11.06 (2.07) | | 12.44 (2.02) | | 14.06 (2.05) | | 13.72 (1.89) | |
| Experimental group | 11.05 (1.97) | (0.998) | 8.60 (1.92) | (0.176) | 6.80 (1.95) | (0.015) * | 3.80 (1.79) | (0.001) ** |
| *Mania scale (Young)* | | | | | | | | |
| Control group | 2.06 (0.68) | | 4.06 (0.81) | | 5.89 (1.17) | | 6.39 (0.95) | |
| Experimental group | 2.50 (0.65) | (0.641) | 0.65 (0.77) | (0.004) * | 0.55 (1.11) | (0.002) * | 0.20 (0.90) | (<0.001) ** |
| *Inadaptation scale* | | | | | | | | |
| Control group | 12.44 (1.63) | | 14.78 (1.34) | | 15.00 (1.29) | | 17.22 (0.90) | |
| Experimental group | 15.10 (1.54) | (0.245) | 4.65 (1.27) | (<0.001) ** | 3.45 (1.22) | (<0.001) ** | 1.90 (0.85) | (<0.001) ** |

* $p<0.05$.
** $p<0.001$.
[a] Bonferroni correction for multiples comparisons.

Figure 3: Isasi et al. (2010) Table 2

One possibility, given the authors are Spanish, is that te stands for "típico error" (typical error), which could be a direct translation of "standard error". I emailed the authors of Isasi et al. (2010) multiple times beginning in September asking for clarification but received no response 10 months later, except for one author who let me know that he is retired and no longer in contact with any of the other coauthors.

Below, I recalculate SMDs from the N, M, and te values reported in Isasi et al. (2010) under the assumption they are a) SDs and b) SEs to assess the plausibility of each.

| timepoint | method | n1i | m1i | sd1i | n2i | m2i | sd2i | smd |
|---|---|---|---|---|---|---|---|---|
| posttreatment | te as SD | 20.00 | 4.06 | 0.81 | 20.00 | 0.65 | 0.77 | 4.32 |
| posttreatment | te as SE | 20.00 | 4.06 | 3.62 | 20.00 | 0.65 | 3.44 | 0.96 |
| follow up | te as SD | 18.00 | 6.39 | 0.95 | 20.00 | 0.20 | 0.90 | 6.70 |
| follow up | te as SE | 18.00 | 6.39 | 4.03 | 20.00 | 0.20 | 4.02 | 1.54 |

- Neither treating the "te"s as SDs nor SEs reproduces the SMDs reported in Gloster et al.'s (2020) forest plots (5.233 for postintervention, 8.02 for follow up), although treating the "te"s as SD more closely approximates the SMDs.
- Treating the "te"s as SEs produces implausibly large SMDs
- Previous studies have estimated the YMRS's SD in clinical samples to be c.4.5, which correspond roughly with treating the "te"s as SEs (N = 211 patients, Targum et al. 2018; N = 209, Suppes et al. 2016).
- Additionally, although not shown here for brevity, Isasi et al. (2010) reported results from other outcome measures, including the BDI which has much better understood distributions: SD in clinical populations are usually 8-12. Plausible BDI SDs were only found when treating the "te"s as SEs.

### Hinton et al. (2005)

Gloster et al. (2020) report the Hinton et al. (2005) effect size as "SMD = 2.22" in their forest plot (Fig 2), and state that they chose the CAPS outcome measure.

I reextracted two of Hinton et al.'s (2005) outcome measures: the CAPS and also the N-PASS. The SMD reported in Hinton et al. (2005) for the CAPS did not match Gloster et al.'s (2020) value, but the value for the N-PASS did.

| outcome | smd | smd_se |
|---|---|---|
| CAPS | 2.17 | 0.40 |
| NPASS | 2.22 | 0.40 |

- The CAPS SMD could not be reproduced.
- However, the SMD value reported in Gloster et al. (2020) does match SMD for the N-PASS (both the value reported in Hinton et al., 2005 and recalculated here). This may be an indication that Gloster et al. (2020) either mislabelled the outcome measure selected or extracted summary statistics for the wrong outcome measure.

### Moore et al. (1997)

Gloster et al. (2020) report the Moore et al. (1997) effect size as "SMD = 1.653" in their forest plot (Fig 2) and state that they extracted data for the HAM-D outcome measure (table 2).

Inspection of Moore et al. (1997) demonstrates that it used an active control condition: it compared Cognitive Therapy against antidepressant medication. A priori, an SMD of 1.65 against an active control is implausibly large.

I reextracted M, SD, and N from Moore et al.'s (1997) for the HAM-D at the posttreatment time point and recalculated SMD:

| smd | smd_se |
|---|---|
| -0.08 | 0.67 |

- Recalculated SMD was close to zero, whereas that reported in Gloster et al. (2020) was very large (SMD = 1.653).

### Watanabe et al. (2011)

I was unable to obtain a copy of this article and could not attempt an effect size reextraction.

## Conclusion

These issues arguably represent clear evidence of major errors that compromise the reliability of the research findings presented in Gloster et al. (2020). Following COPE guidelines, the work requires substantial correction or possibly retraction (COPE, 2019).

This assessment of the trustworthiness of the results presented in Gloster et al. (2020) is not exhaustive and other issues may be present.

No attempt was made to correct these issues or consider how the conclusions of Gloster et al. (2020) would be affected or adjusted if these and any other issues present were fixed. Requiring those who highlight serious issues with published work to also solve those issues creates perverse incentives in science, as it moves the burden of proof from original authors to critics (see Hussey, 2025).

## Data and code availability

All data and code available at https://github.com/ianhussey/verification-gloster-et-al-2020