# Problematic claims about the superiority of

# hierarchical versus distinction relations in a self-based ACT exercise:

# A reanalysis of Foody, Barnes-Holmes, Barnes-Holmes & Luciano (2013)

# Ian Hussey

The evidential link between Acceptance and Commitment Therapy (ACT) and Relational Frame Theory (RFT) is a matter of on-going debate. Foody et al. (2013) represents a well-cited study on what they argue is an RFT analysis of what ACT refers to as self-as-context. In the spirit of the behavioral community’s recent embrace of the issue of the credibility and replicability of our field’s claims, I provide a critique and reanalysis of Foody et al. (2013). I found many issues: the authors’ results are not robust to appropriate corrections; their claims do not correspond with their analyses; their results are seriously confounded by regression to the mean; this confound implies that their distinction intervention may actually have prolonged participants’ distress; and results generally did not replicate in the original authors’ own unpublished direct self-replication study. Taken together, results illustrate how flawed research practices can give rise to claims with direct clinical implications that, on inspection, lack credibility and replicability, and may possibly even cause harm.

Foody, Barnes-Holmes, Barnes-Holmes, and Luciano (2013) reported the results of a study in which they compare the efficacy of two different interventions, both based on elements of Acceptance and Commitment Therapy (ACT) (Hayes et al., 1999), in relieving experimentally induced distress. Foody et al.’s (2013) study presented itself as a conceptual replication of a previous study by Luciano et al. (2011), both of which attempt to provide a more technical analysis of behavioral processes, from the perspective of Relational Frame Theory (RFT: Hayes et al., 2001), of the behavioral processes involved in what ACT refers to as defusion and self-as-context (Hayes et al., 1999). Foody et al. (2013) stated that they and the study by Luciano et al. (2011) were among the first to successfully attempt to bridge the gap between ACT and RFT, specifically by recasting what ACT calls the therapeutic processes of “defusion” and “the three selves” into the more precise language of what RFT refers to as deictic relational responding (see Foody et al., 2013, Luciano et al. 2011). Specifically, Foody et al. (2013) adapted two different ultra-brief therapeutic interventions designed to decrease distress from Luciano et al. (2011). Briefly, these defined the distinction condition, in which participants were instructed to attempt to see their thoughts and feelings as distinct from their sense of self (i.e., you are not your thoughts), versus the hierarchy condition, in which participants were instructed to attempt to see their thoughts and feelings as contained by an overarching self of self (i.e., you contain your thoughts). These two studies therefore represent key papers in the on-going debate about the strength of the evidence for ACT’s core processes and their ties to basic science via RFT (see Barnes-Holmes et al., 2015; McLoughlin & Roche, 2022). Both are well cited, with 139 and 112 citations, respectively, on Google Scholar at time of writing.

Elsewhere, over the past decade, the Replication Crisis in psychology has raised questions about the replicability, robustness, and credibility of claims in the psychology literature (Gelman, 2016; Spellman, 2015). Although the replication crisis began in social psychology, recognition of the same systemic weaknesses, flaws and biases in our research processes have more recently also been acknowledged in clinical psychology (Leichsenring et al., 2017; Tackett et al., 2019). More recently, awareness of this issue has also spread to the behavioral research communities. In an editorial for Perspectives on Behavior Science, Hantula stated that “the ‘replication crisis’ in psychology could well be repeated in behavior science and behavior analysis. Even if it is not, it may hold some important lessons for both scientists and practitioners.”(Hantula, 2019, pp. 4-5). Encouragingly, the Association for Contextual Behavioral Science’s Task force on the Strategies and Tactics of Contextual Behavioral Science Research (2021) recently announced its explicit support for Open Science principles, including data transparency and a focus on replication. As such, there is now growing support for the idea that behavioral research would be enhanced by examining and enhancing the reproducibility and credibility of its claims. As such, it seems important to reexamine the results and claims presented in Foody et al. (2013).

## Summary of Foody et al. (2013)

### Stated relevance

Foody et al. (2013) stated the relevance of their study as follows: “The current study is among the first to attempt to target specific relational frames in the context of ACT exercises. In doing so, it fits the broader research agenda of scientific bridge building between ACT and RFT, while recognizing the difficulties inherent in the use of middle level terms, such as self as context and defusion. One of the central ways forward in dealing with middle level terms is to replace them with more functionally sound, empirically tested concepts, such as replacing the terms self as context with distinction or hierarchical deictic relations. Although the present study is only one small step in that direction, it does suggest that RFT concepts may have more clinical application than might have been previously recognized.” (Foody et al., 2013, p. 387).

### Design and method

Foody et al. (2013) employed a 3 (within time points: baseline, post distress induction, post ACT intervention) X 2 (between intervention groups: “hierarchical self as context” vs. “distinction self as context”) mixed between-within design. Three primary outcome measures were assessed at each time point: three single-item visual analogue scales (VAS) “were used as distress ratings and assessed discomfort, anxiety, and stress” (Foody et al., 2013, p. 376). Each visual analogue scale required participant to indicated “their level of distress on each scale by placing an X on a printed line that ranged from 0% (e.g., no discomfort) to 100% (e.g., very much discomfort).” (Foody et al., 2013, p. 376). Secondary outcome measures will not be considered here for brevity. Participants were assessed at baseline, completed a distress induction task, then were assessed again (post induction), then completed an ACT intervention (randomised to either a “hierarchical self as context” or “distinction self as context” exercise), and then completed the assessments again (post intervention). The analyses included 18 participants per group after exclusions.

### Hypothesis and claims

Foody et al. (2013) compared the efficacy of two interventions in relieving experimentally-induced distress. Their stated claim was that the “hierarchical self as context” intervention as more effective than the “distinction self as context” intervention, and they state that they therefore conceptually replicated the results of the original study by Luciano et al. (2011). In their own words, they summarize their key findings as follows: “The findings demonstrated superiority of the intervention that focused on hierarchical, rather than distinction, deictic relations in terms of reducing distress.” (p. 373); “The superiority observed for the hierarchical intervention, relative to distinction, bore some overlap with the findings from the original study.” (p. 384); “The hierarchical intervention only resulted in a reduction in all three dependent measures, including a significant reduction in stress. … The lack of effect for the distinction intervention is also similar to the findings from the original [study], in which Luciano et al. found only limited effects for the defusion I intervention.” (p. 385); and “the hierarchical intervention was significantly effective only in the context of stress, and not in discomfort or anxiety (although both of these were also reduced).” (p. 385).

## Critiques

### Results are not robust to appropriate corrections for multiple comparisons

In their abstract, Foody et al. (2013) stated that their “findings demonstrated superiority of the intervention that focused on hierarchical, rather than distinction, deictic relations in terms of reducing distress.” (p. 373). They reiterated this claim in the first paragraph of their discussion: “Nonetheless, the superiority observed for the hierarchical intervention, relative to distinction, bore some overlap with the findings from the original study.” (p. 384). However, a few paragraphs later they stated that the effect “was significantly effective only in the context of stress, and not in discomfort or anxiety” (p. 385), that is, for only one of the three outcome measures. This means that the authors made conclusions about the differential efficacy in general (i.e., on “distress”) based on just one of three measures of distress showing a significant effect.

Unfortunately, the authors did not explicitly tie these conclusions to the results of specific statistical inference tests. We can assume that these claims were based on the statistical significance of the interaction effects in three mixed within-between RM-ANOVAs that employed the outcome measures as dependent variables (in separate models), time point as within-subjects independent variable, and condition as between-groups independent variable (p. 381-382), as no other set of results in the article followed this pattern or has the same relevance to the claim. They reported that the interaction effects were significant for stress (“*p* = .04”) but not discomfort (“*p* = .45”) or anxiety (“*p* = .33”; pp. 381-382).

At this point, it is useful to note that this combination of barely-significant *p* values, small sample sizes, multiple outcome measures, and global conclusions being made on the basis of a subset of statistically significant results represent indications that the results may not be credible (REFs).

However, let us take these analyses at face value for a moment. Based on the above quotes, Foody et al.’s (2013) inferential method can be summarized as accepting their alternative hypothesis if they obtained significant results on any of the three outcome variables. Given this inference method, good statistical practice would require that these results are corrected for the familywise error rate (REF). Simply put, if one is willing to accept the alternative hypothesis on the basis of any significant result across multiple outcome measures, alpha corrections must be applied in order to keep the long run false positive rate within the nominal alpha value (e.g., 5%). Luckily, these corrections can be applied post hoc using the reported *p* values. Applying even a liberal correction method (e.g., Holm corrections, implemented using R’s p.adjust function) produces three non-significant adjusted *p* values (i.e., discomfort: *p*adj = .66, anxiety: *p*adj = .66, stress: *p*adj = .12). Using the results of their own statistical models with appropriate alpha corrections applied to their results, Foody et al.’s (2013) results therefore do not support their conclusion that the hierarchy intervention more effectively relieves distress than the distinction intervention.

### Claims were not linked to the actual analyses reported

However, one could argue that even adjusted *p* values from the RM-ANOVAs are also uninformative to the actual claim that the hierarchical condition is superior to the distinction condition: by including the baseline scores, the interaction effects do not actually test the hypothesis that the interventions produce differential effects, because interaction effects could be driven by one or more of the baseline, post induction and post intervention time points. In order to support their claim, the authors would need to have reported the results of post hoc contrasts exploring these interaction effects, or to compare the post intervention conditions directly via a *t*-test. However, Foody et al. (2013) do not report any such results. Foody et al. (2013) do present mean change scores between time points for each of the visual analogue scales: “distinction resulted in a very small increase in discomfort (+.76), while hierarchy resulted in a decrease (-7.57)”; “Anxiety subsequently decreased for both conditions, although the larger change was recorded for the hierarchical intervention (distinction: -.03; hierarchy: -3.86)”; and “distinction resulted in an increase in stress (+4.71), while hierarchy reduced stress (-8.82).” (pp. 381-382). However, inferences about the population effect cannot be made on the basis of the sample means alone. As such, in summary, Foody et al. (2013) suffers from an absence of appropriate analyses to test their stated claim that the hierarchical condition is superior to the distinction condition.

There is good reason to believe that Foody et al. were aware that some form of pairwise comparisons between groups or timepoints beyond the RM-ANOVAs would be informative: because Foody, in her unpublished direct replication of their 2013 study (Foody, 2013, experiment 10) reported such pairwise comparisons between timepoints for each condition. – and claimed she found significant results where she did not (e.g., p.180).

### Claims are strongly confounded by regression to the mean

Of course, one might argue that the RM-ANOVAs serve a different but possibly useful purpose: perhaps they help us understand changes between time points. Unfortunately, there are strong reasons to believe that the study design precludes meaningful interpretations here either due to a serious confound. A confound is a variable that systematically covaries with the independent variable (i.e., the impact of the interventions on distress across time points) and affects the dependent variable (i.e., responses on the visual analogue scales), but is not intentionally manipulated by the researcher. Confounding variables can obscure the true effects of the independent variable on the dependent variable, and lead to inaccurate or misleading conclusions.

Specifically, Foody et al.’s (2013) design is confounded by a phenomenon called regression to the mean. This refers to the tendency for extreme scores on a measure to move closer to the average score over time with repeated measurements. Participants who were induced with distress by the stressful task were likely to have high scores on distress before the intervention. These high scores were likely to regress towards their average level of distress over time, regardless of which intervention they received. Therefore, it is possible that some or all of the observed reduction in distress after the intervention is due to regression to the mean rather than the effect of the intervention. To avoid this confound, the authors should have, for example, included a negative control group that did not receive any intervention after the distress induction procedure. Differences between the negative control condition and the two intervention conditions could have then been compared in order to test the claim that it was the interventions specifically that decreased distress rather than merely natural decrease in distress over time.

There is very good reason to think that Foody et al. (2013) should have been aware of this confound and therefore included such a negative control condition: this confound is not conjecture, but has been repeatedly observed in previous studies that Foody et al. themselves cited in their previously published article on the efficacy of the distress induction procedure (Foody et al., 2012). In that publication, Foody et al. (2012) stated that the single-sentence distress induction procedure was created by Rachman et al. (1996) and cite four other previous publications that employed the same procedure: van den Hout et al. (2002), Bocci & Gordon (2007), Marcks & Woods (2007), and Zucker et al. (2002).

Inspection of those articles demonstrated that two of them, including the original publication, reported anxiety scores from a visual analogue scale, employed a no-instructions negative control condition, and reported data from a follow-up time period (Rachman et al., 1996; van den Hout et al., 2002). Note that whereas Foody et al. (2013) refer to this third time point as post intervention, it is referred to as post delay in the control condition as no intervention was used. Both of those publications present results demonstrating that distress increased after the induction procedure, but then fell sharply again after a short delay (e.g., two minutes) and in the absence of any intervention.

In order to illustrate this effect, I reanalyzed data from those two studies. I extracted summary statistics (sample sizes, means and standard deviations) for these negative control conditions. Summary statistics were then converted to Hedges’ *g* standardized effect sizes, their 95% Confidence Intervals, and Welch’s independent *t*-tests (see method section for equations). This was done as if the data were fully between groups (i.e., the best available approximation that was possible with the limited data reported in the articles).

Results from the plotted means, the *t*-tests, and the effect sizes and their confidence intervals collectively describe the same pattern of effect in both studies: the distress induction procedure successfully increased anxiety (i.e., large increases between baseline and post induction), and after a short delay anxiety naturally decreased back to near-baseline levels in the absence of any intervention (i.e., large decreases between post induction and post delay; no significant differences between baseline and post intervention). These effects are illustrated in Figure 1 and results can be found in Table 1.

In summary, Foody et al. (2013) state that the reductions in distress were due to the interventions. They fail to acknowledge the potential for regression to the mean, which confounds inferences about changes between the post induction and post intervention time points being due to the interventions. Foody et al. (2013) did not include a negative control condition that could have allowed them to control for this confound, despite the fact that the literature on this distress induction procedure – which they themselves cite in a previous publication examining the induction procedure (Foody et al., 2012) – has routinely done so. Such studies have demonstrated that, within a few minutes, distress decreases from the levels observed at post induction even in the absence of any intervention. Foody et al.’s (2013) observed changes in distress between post induction and post intervention therefore cannot clearly be attributed to the intervention.

### The distinction condition may have actually prolonged distress relative to doing nothing

An alternative interpretation of Foody et al.’s (2013) results that is both plausible and worrying is that the distinction intervention prolonged distress compared to what would have been observed in the absence of any intervention. Put simply, if the natural course of distress within this paradigm is to reduce within a short period of time, and Foody et al. (2013) observed distress staying the same or worsening in their distinction condition, then it is plausible that the intervention is worse than doing nothing.

Foody et al. (2013) did not include a negative control condition and so direct comparisons cannot be made, nor can this confound be directly statistically controlled for. However, it is useful to nonetheless consider the descriptive trends observed in Foody et al. (2013) and compare them with those observed in the negative control conditions in previous publications.

As discussed previously, published studies shown cited by Foody et al. have shown that anxiety decreases sharply from post induction to post delay time points in the absence of any intervention. Indeed, anxiety returns to near-baseline levels (see Figure 1). This is in stark contrast to the results of Foody et al.’s (2013) distinction intervention condition, who reported that sample means for anxiety, discomfort, and stress either stayed roughly the same or increased between post induction and post intervention (*M*diff = -0.03*,* +0.76, +4.71 respectively). Foody et al. (2013) concluded that only the hierarchy condition was effective in relieving distress (p. 385), and by implication that the distinction condition was therefore ineffective. However, when descriptively contrasted with the control conditions observed in previous work (Rachman et al., 1996; van den Hout et al., 2002), it may be more accurate to have concluded that the distinction condition serves to prolong or maintain distress at the post intervention time point. The potential that the interventions may prolong distress rather than relieve it underscores the need to employ negative control conditions, especially in the face of already known confounds such as regression to the mean.

**Figure 1.** Results from no-instruction negative control conditions reported in previous studies using the distress induction procedure which were cited in Foody et al. (2012). Points represent means, error bars represent 95% Confidence Intervals.

## 

**Table 1.** Summary statistics, effect sizes, and results of Welch’s independent *t*-tests from no-instruction negative control conditions reported in previous studies using the distress induction procedure which were cited in Foody et al. (2012).

|  |  |  | 95% CI | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Study | Comparison | Hedges’ *g* | Lower | Upper | *t* | df | *p* |
| Rachman et al. (1996) | Baseline - post induction | -4.69 | -5.61 | -3.76 | -19.56 | 56.45 | < .00001 |
| Rachman et al. (1996) | Post induction - post delay | 3.24 | 2.51 | 3.97 | 13.53 | 62.25 | < .00001 |
| Rachman et al. (1996) | Baseline - post delay | -0.21 | -0.68 | 0.27 | -0.87 | 48.68 | .38849 |
| van den Hout et al. (2002) | Baseline - post induction | -1.46 | -1.95 | -0.96 | -6.58 | 44.22 | < .00001 |
| van den Hout et al. (2002) | Post induction - post delay | 1.17 | 0.69 | 1.64 | 5.28 | 56.16 | < .00001 |
| van den Hout et al. (2002) | Baseline - post delay | -0.38 | -0.82 | 0.06 | -1.71 | 59.86 | .09182 |

# Method

Foody et al. (2013) does not report a direct, appropriate and unconfounded statistical inference test to support their core claim that the hierarchy condition was produced lower levels of distress than the distinction condition. I therefore conducted a reanalysis to test this claim. Specifically, I aimed compare distress between the two groups at the post intervention time point using independent Welch’s *t*-tests with Hedges’ *g* standardized effect sizes (i.e., Cohen’s *d* with correction for small sample sizes). However, it should be noted that even this reanalysis cannot control for the confound noted previously. As such, these analyses quantify the evidence for the specific claim regarding the superiority of the hierarchy intervention relative to the distinction intervention, rather than the superiority of either or both relative to doing nothing. All raw and processed data as well as R code for data processing and analyses are available (osf.io/XXX).

## Attempts to obtain additional information from the original authors

In the first instance, I attempted to obtain the original dataset. I contacted all authors of Foody et al. (2013) asking if they’d be willing to share the original data. The first author informed me that the dataset no longer exists. I also asked the original authors whether the error bars in Foody et al.’s (2013) Figures 1 to 3 represent 95% Confidence Intervals or Standard Error of the Mean, as this wasn’t reported in the manuscript. However, the authors were unresponsive to subsequent requests.

## Extraction of summary statistics

Independent Welch’s *t*-tests and Hedges’ *g* effect sizes were constructed from summary statistics without access to the raw data, specifically from the sample size (*n*), mean (*M*) and standard deviation (*SD*) for each condition and outcome variable at each time point. Sample sizes after exclusions for both conditions were reported in text: “Participants were allocated randomly across two conditions denoted as distinction self as context (N= 18) and hierarchical self as context (N= 18).” (p. 375).

Means for each time point were not reported in text, only approximate values for the baseline time point (e.g., “<11”). However, (a) change scores for both conditions between the time points were reported in text (pp. 381-383) and (b) means were plotted in Figures 1 to 3. I therefore extracted estimates of the means for each condition and time point from the plots. Means for the post intervention time point were calculated in two different way to validate them against one another: using the mean for that time point extracted from the plots; and using the mean for the baseline time point adding the change scores between time points reported in text. Both results produced estimate that were all less than ±0.6 (on a 0 to 100 scale), suggesting that the extracted estimates are very close to the values used to generate the plots. Given their extremely high similarity and the fewer number of steps involved in the latter (and therefore fewer opportunities for errors to be introduced, for example via rounding), I employed estimates obtained via just the latter method for the below analyses.

Standard deviations can be recalculated from both Confidence Intervals and Standard Errors of the Mean. However, the conversion formula to do this depends on the type of the interval that was reported. Foody et al. (2013) did not report in their manuscript what the intervals in their Figures 1 to 3 represent, and the authors did not reply to my questions about this via email. In the absence of an answer, I calculated both: two sets of Standard Deviations and analyses based on them, one assuming that they are 95% CIs and one assuming they are SEMs, with the intention of considering the plausibility of both sets of results. The width of the intervals for both conditions in the post intervention time point were extracted from Foody et al.’s (2013) Figures 1 to 3. No intervals were reported numerically in their text. These intervals were converted to standard deviations using both of the below equations, one assuming they are 95% Confidence Intervals and one assuming that they are the Standard Error of the Mean. The extracted means and recalculated standard deviations for each outcome variable in both conditions at the post intervention time point can be found in Table 2.

# Results

## Reanalysis of Foody et al. (2013)

### Differences between the conditions for each outcome variable

The means, standard deviations, and sample sizes were then used to calculate independent Welch’s *t*-tests. This was done by calculating the Standard Error of difference in means (SE), *t* value (*t*), and degrees of freedom (df) using the below equations.

Hedge’s *g* effect size, a version of Cohen’s *d* with a bias correction for small sample sizes, was calculated using the following equation.

*p* values were calculated from the estimates of *t* and df using the *t* distribution. Finally, given the multiple correlated outcomes corresponding to the same underlying hypothesis, adjusted *p* values (*p*adj) were calculated in order to correct for the familywise error rate using Holm corrections (see Table 2 and Figure 2).

#### Assuming intervals represent 95% CIs

If we assume the intervals reported in Foody et al.’s (2013) plots represent 95% CIs, results demonstrated that scores were significantly lower in the hierarchical condition than the distinction condition for discomfort (*p* = .005, *p*adj = .014) and stress (*p* = .009, *p*adj = .018), but no statistically significant decreases were observed for anxiety (*p* = .840, *p*adj = .840; see Table 2 and Figure 2 [upper panel, blue bars]).

However, the effect size for discomfort and stress were so large as to possibly raise questions about their credibility (Hedges’ *g* = 0.99 and 0.91, respectively). This is especially the case when one considers that these effect sizes represent not merely the impact of the intervention but the differential impact between the two interventions (for discussion of the plausibility of very large effect sizes see Funder & Ozer, 2019; Hilgard, 2021).

In order to provide a meaningful comparison,

**Table 2.** Effect sizes for the distress induction intervention

|  |  |  |  |
| --- | --- | --- | --- |
|  | Difference in means | |  |
| Outcome | Change due to distress induction | Differences between interventions at post intervention | Percent (comparison/induction) |
| Anxiety | 1.50 | 0.07 | 5% |
| Discomfort | 0.92 | 0.51 | 55% |
| Stress | 0.51 | 0.46 | 90% |

One possible explanation for these very large effect sizes would be that one condition actually served to prolong discomfort and stress relative to doing nothing, as discussed previously in the introduction.

#### Assuming intervals represent the SEM

If we assume the intervals reported in Foody et al.’s (2013) plots represent the SEM, no significant differences were found between the conditions on any of the outcome variables, whether considering either native or adjusted *p* values: discomfort (*p* = .131, *p*adj =.392), stress (*p* = .918, *p*adj =.918), and anxiety (*p* = .166, *p*adj =.392; see Table 2 and Figure 2 [lower panel, blue bars]).

**Table 2.** Independent Welch’s *t*-tests comparing the hierarchical and distinction conditions at the post intervention time point. Adjusted *p* values using Holm corrections. Results calculated assuming that the intervals reported in Foody et al.’s (2013) plots represent 95% Confidence Intervals or the Standard Error of the Mean.

|  |  | Distinction | | | Hierarchy | | |  | 95% CI | |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Interval | DV | *M* | *SD* | *n* | *M* | *SD* | *n* | Hedges’ *g* | Lower | Upper | *t* | df | *p* | *p*adj |
| 95% CI | Discomfort | 24.76 | 8.66 | 18 | 15.43 | 9.74 | 18 | 0.99 | 0.29 | 1.68 | 3.04 | 33.54 | .005 | .014 |
|  | Anxiety | 17.97 | 12.18 | 18 | 17.14 | 12.18 | 18 | 0.07 | -0.59 | 0.72 | 0.20 | 34.00 | .840 | .840 |
|  | Stress | 21.71 | 10.82 | 18 | 12.18 | 9.74 | 18 | 0.91 | 0.21 | 1.59 | 2.78 | 33.63 | .009 | .018 |
| SEM | Discomfort | 24.76 | 16.97 | 18 | 15.43 | 19.09 | 18 | 0.51 | -0.16 | 1.17 | 1.55 | 33.54 | .131 | .392 |
|  | Anxiety | 17.97 | 23.86 | 18 | 17.14 | 23.86 | 18 | 0.03 | -0.62 | 0.69 | 0.10 | 34.00 | .918 | .918 |
|  | Stress | 21.71 | 21.21 | 18 | 12.18 | 19.09 | 18 | 0.46 | -0.20 | 1.12 | 1.42 | 33.63 | .166 | .392 |

**Figure 2.** Hedges’ *g* effect sizes for each distress outcome variable reported in Foody et al. (2013) (in blue) plus a pooled variable (in green), calculated by treating the intervals reported in Foody et al. (2013) as either 95% Confidence Intervals or Standard Errors of the Mean.



#### Differences between outcomes assuming intervals are 95% CIs

If we assume that the intervals reported in Foody et al.’s (2013) plots represent 95% CIs, results present an unclear picture: statistically significant differences were found between the conditions on discomfort and stress but not anxiety. It is important to appreciate that this does not mean that differential effects were observed between the outcome variables, as the difference between significant and non-significant is not necessarily itself significant (Gelman & Stern, 2006). Put another way, if we want to conclude that that the interventions produced differential outcomes between the three outcome measures, this would need to be tested directly. I did this by converting the Hedges’ *g* effect sizes and their confidence intervals for each outcome measure to pairwise *Z* scores using the below equation. *Z* scores were then converted to *p* values via the normal distribution.

No differences were found in any of the pairwise comparisons: discomfort vs. anxiety, *p* = .361; discomfort vs. stress, *p* = .999; anxiety vs. stress *p* = .361. Foody et al.’s (2013) results therefore cannot be interpreted as evidence of differential impact of the interventions between the three outcome variables.

### Differences between the conditions on the pooled outcomes

As discussed previously, Foody et al. (2013) make claims in their abstract and discussion regarding the superiority of the hierarchical intervention over the distinction intervention with regard to decreasing “distress” in general rather than their three component outcome measures (discomfort, anxiety, and stress). Whether or not it is appropriate to treat these three ad hoc measures as valid measures of a latent “distress” variable cannot be answered based on the summary statistics alone and would require separate prior measure validation. However, putting this measurement question aside and following Foody et al.’s (2013) own logic, it is possible to calculate a single statistical inference test to test their core claim regarding distress as a whole. This would also serve to resolve the previously stated issue of (a) a mix of significant and non-significant results between outcome measures and yet (b) no evidence of significant differences between the measures. In order to provide this single test of Foody et al.’s (2013) claim regarding distress as a whole, I calculated pooled means for each condition by averaging them, and pooled standard deviations via the following formula.

These were used to calculate a further set of Welch’s t-tests and effect sizes (i.e., one assuming the intervals were 95% CIs and one assuming SEMs). Regardless of whether the intervals reported in Foody et al.’s (2013) plots represent 95% CIs or SEMs, no significant differences were found between the conditions (*p*s = .073, .352 respectively). Results can be found in in Table 3 and are illustrated in Figure 2 (green bars).

**Table 3.** Independent Welch’s *t*-test comparing the hierarchical and distinction conditions at the post intervention time point using the pooled responses on the visual analogue scales.

|  | Distinction | | Hierarchy | |  | 95% CI | |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Interval | *M* | *SD* | *M* | *SD* | Hedges’ *g* | Lower | Upper | *t* | df | *p* |
| 95% CI | 21.48 | 10.65 | 14.92 | 10.61 | 0.60 | -0.07 | 1.27 | 1.85 | 34 | .073 |
| SEM | 21.48 | 20.88 | 14.92 | 20.80 | 0.31 | -0.35 | 0.96 | 0.94 | 34 | .352 |

## Reanalysis of Foody’s (2013) unpublished direct replication (thesis experiment 10)

Ultimately,

"One single observation may be the stray result of some trivial error: in order to reduce such risks, methodological falsificationists prescribe some safety control. The simplest such control is to repeat the experiment" Lakatos, 1978

## Assessment of whether intervals were likely to be 95% CIs or SEMs

Given that several of the previously reported reanalyses provide different results depending on whether we assume the intervals reported in Foody et al. (2013) were 95% CIs or SEMs, and the authors of Foody et al. (2013) are unable or unwilling to answer this question, it seems important to gather relevant information to aid our abduction here.

One useful source of information is what the authors’ of Foody et al. (2013) report in their other publications. Publications that came from the Barnes-Holmes research group around the time as Foody’s articles were published employed SEMs as error bars in their plots (e.g., Hughes & Barnes-Holmes, 2011). This includes one of my own studies which, I should note, also failed to clarify this in text (Hussey & Barnes-Holmes, 2012). My personal recollection is that we were generally trained by them to plot SEMs. This idea is supported by the fact that some of Foody’s recent behavioral publications have employed SEMs (e.g., Maloney et al., 2020).

Another source of information is to compare the known standard deviations reported in previously published studies with the standard deviations extracted from Foody’s work under both assumptions of what the intervals represent (Foody et al., 2012; Foody et al., 2013; Foody, 2013). Given that the interventions involve an unknown degree of inter-individual heterogeneity, it is safer to compare the standard deviations of the baseline and post induction time points between studies.

In order to divine what the intervals represented,

# Discussion

“We need less research, better research, and research done for the right reasons” (Altman, 1994)

“Most scientific studies are wrong, and they are wrong because scientists are interested in funding and careers rather than truth.” (Ioannidis, 2005)

Reanalyses of Foody et al.’s (2013) results suggest that their data do not represent credible evidence of differences between the intervention conditions.

Possible explanations for Foody et al.’s (2013) results.

* Error variance, i.e., where it did find significant results, these were false positives. They don’t replicate the effects in their own unpublished direct replication study, although they did choose to use a smaller sample size (12 instead of 18 per group).

**Figure XX.** Decision tree for how the results of Foody et al. (2013) and Foody’s unpublished direct replication (thesis experiment 10) should be interpreted based on different assumptions and analytic decisions.



Statistically significant results of differences between the conditions is only found under a limited range of exceptionally liberal assumptions, all of which would need to be met: First, the intervals they reported in the plots must would need to represent 95% CIs rather than SEMs, given that results computed from the latter are all null. Second, one would need to disagree that alpha corrections were appropriate despite the fact that the three outcome variables (anxiety, discomfort, and stress) were used to make a single conclusion (i.e., regarding “distress”), given that results computed with alpha corrections are all null. Third, one would need to additionally disagree that it is appropriate to pool the outcome measures to test differences in distress more directly, given that results computed for the pooled outcome variable are null. Fourth, one would need to accept that extremely large effects sizes (i.e., Hedges’ *g* ≥ .91) between two active treatment groups were plausible (for discussion of the plausibility of very large effect sizes see Funder & Ozer, 2019; Hilgard, 2021). Fifth and finally, one would have to disregard the results of Foody et al.’s own unpublished direct self-replication, in two senses: (a) it may represent evidence of publication bias given that no significant results were found following the original authors’ analytic strategy; and second, and also (b) that upon reanalysis statistically significant results were now found for only one of the three outcome variables (and even then, only under assumptions 1, 2, 3, and 4 above).

The alternative inference from Foody et al.’s (2013) results, should one not agree with one or more of the above, would be that they do not represent credible evidence of differences between the distinction and hierarchy intervention conditions.

If Foody et al.’s (2013) results are actually null, this also has implications for Luciano et al. (2011), which Foody et al. (2013) state they are conceptually replicating.

Other work attempting to tie RFT principles to ACT practices, such as the use of metaphor in therapy, have also been presented (Sierra et al., 2016) but have also failed to replicate (Pendrous et al., 2020; for a series of replies, in order of publication, see: Hulbert-Williams et al., 2020; Ruiz et al., 2020; Hussey, 2020).

It is a shame that the disconnect between Foody et al.’s (2013) results and claims was not caught during the peer review process.

References

Barnes-Holmes, Y., Hussey, I., McEnteggart, C., Barnes-Holmes, D., & Foody, M. (2015). Scientific ambition: The relationship between Relational Frame Theory and middle-level terms in acceptance and commitment therapy. In R. D. Zettle, S. C. Hayes, D. Barnes-Holmes, & A. Biglan (Eds.), *The Wiley Handbook of Contextual Behavioral Science* (pp. 365–382). Blackwell-Wiley. http://onlinelibrary.wiley.com/doi/10.1002/9781118489857.ch18/

Bocci, L., & Gordon, P. K. (2007). Does magical thinking produce neutralising behaviour? An experimental investigation. *Behaviour Research and Therapy*, *45*, 1823–1833. https://doi.org/10.1016/j.brat.2007.02.003

Curtis, A., Foody, M., Kelly, M., & Murphy, C. (2020). Using the Implicit Relational Assessment Procedure to explore students’ implicit age discrimination of toxic versus innocuous phrases. *International Journal of Psychology & Psychological Therapy*, *20*(3), 241–257. APA PsycInfo.

Fleming, K., Foody, M., & Murphy, C. (2020). Using the Implicit Relational Assessment Procedure (IRAP) to examine implicit gender stereotypes in science, technology, engineering and maths (STEM). *The Psychological Record*, *70*(3), 459–469. APA PsycInfo. https://doi.org/10.1007/s40732-020-00401-6

Foody, M. (2013). *An Empirical Investigation of Self: Bridging the Gap between ACT, Mindfulness and RFT* [National University of Ireland Maynooth]. http://eprints.maynoothuniversity.ie/4777/

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gelman, A. (2016, September 21). What has happened down here is the winds have changed. *Statistical Modeling, Causal Inference, and Social Science*. http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/

Gelman, A., & Stern, H. (2006). The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Hantula, D. A. (2019). Editorial: Replication and Reliability in Behavior Science and Behavior Analysis: A Call for a Conversation. *Perspectives on Behavior Science*, *42*(1), 1–11. https://doi.org/10.1007/s40614-019-00194-2

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. Kluwer Academic/Plenum Press.

Hayes, S. C., Strosahl, K., & Wilson, K. G. (1999). *Acceptance and Commitment Therapy: An experiential approach to behavior change*. Guilford Press.

Hilgard, J. (2021). Maximal positive controls: A method for estimating the largest plausible effect size. *Journal of Experimental Social Psychology*, *93*, 104082. https://doi.org/10.1016/j.jesp.2020.104082

Hughes, S., & Barnes-Holmes, D. (2011). On the formation and persistence of implicit attitudes: New evidence from the Implicit Relational Assessment Procedure (IRAP). *The Psychological Record*, *61*(3), 391–410. psyh. https://doi.org/10.1007/BF03395768

Hulbert-Williams, L., Pendrous, R., Hochard, K. D., & Hulbert-Williams, N. J. (2020). In search of scope: A response to Ruiz et al. (2020). *Journal of Contextual Behavioral Science*. https://doi.org/10.1016/j.jcbs.2020.10.008

Hussey, I. (2020). *General claims require generalized effects: A reply to Ruiz et al.’s (2020) ‘A systematic and critical response to Pendrous et al. (2020) replication study.’* PsyArXiv. https://doi.org/10.31234/osf.io/83z2y

Hussey, I., & Barnes-Holmes, D. (2012). The Implicit Relational Assessment Procedure as a measure of implicit depression and the role of psychological flexibility. *Cognitive and Behavioral Practice*, *19*(4), 573–582. psyh. https://doi.org/10.1016/j.cbpra.2012.03.002

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Leweke, F., Luyten, P., Keefe, J. R., Midgley, N., Rabung, S., Salzer, S., & Steinert, C. (2017). Biases in research: Risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychological Medicine*, *47*(6), 1000–1011. https://doi.org/10.1017/S003329171600324X

Luciano, C., Ruiz, F. J., Torres, R. M. V., & Mar, V. S. (2011). A Relational Frame Analysis of Defusion Interactions in Acceptance and Commitment Therapy. A Preliminary and Quasi-Experimental Study with At-Risk Adolescents. *International Journal of Psychology*.

Maloney, E., Foody, M., & Murphy, C. (2020). Do response options in the Implicit Relational Assessment Procedure (IRAP) matter? A comparison of contextual relations versus relational coherent indicators. *The Psychological Record*, *70*(2), 205–214. APA PsycInfo. https://doi.org/10.1007/s40732-019-00360-7

Marcks, B. A., & Woods, D. W. (2007). Role of thought-related beliefs and coping strategies in the escalation of intrusive thoughts: An analog to obsessive-compulsive disorder. *Behaviour Research and Therapy*, *45*(11), 2640–2651. https://doi.org/10.1016/j.brat.2007.06.012

McLoughlin, S., & Roche, B. T. (2022). ACT: A Process-Based Therapy in search of a process. *Behavior Therapy*, S0005789422001022. https://doi.org/10.1016/j.beth.2022.07.010

Pendrous, R., Hulbert-Williams, L., Hochard, K. D., & Hulbert-Williams, N. J. (2020). Appetitive augmental functions and common physical properties in a pain-tolerance metaphor: An extended replication. *Journal of Contextual Behavioral Science*, *16*, 17–24. https://doi.org/10.1016/j.jcbs.2020.02.003

Rachman, S., Shafran, R., Mitchell, D., Trant, J., & Teachman, B. (1996). How to remain neutral: An experimental analysis of neutralization. *Behaviour Research and Therapy*, *34*(11–12), 889–898. https://doi.org/10.1016/S0005-7967(96)00051-4

Ruiz, F. J., Luciano, C., & Sierra, M. A. (2020). A systematic and critical response to Pendrous et al. (2020) replication study. *Journal of Contextual Behavioral Science*. https://doi.org/10.1016/j.jcbs.2020.04.011

Sierra, M. A., Ruiz, F. J., Flórez, C. L., Hernández, D. R., & Luciano, C. (2016). The Role of Common Physical Properties and Augmental Functions in Metaphor Effect. *International Journal of Psychology*, 15.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. https://doi.org/10.1177/1745691615609918

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology’s Replication Crisis and Clinical Psychological Science. *Annual Review of Clinical Psychology*, *15*(1), 579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710

Task Force on the Strategies and Tactics of Contextual Behavioral Science Research. (2021). *Adoption of Open Science Recommendations | Association for Contextual Behavioral Science*. https://contextualscience.org/news/adoption\_of\_open\_science\_recommendations

van den Hout, M., Kindt, M., Weiland, T., & Peters, M. (2002). Instructed neutralization, spontaneous neutralization and prevented neutralization after an obsession-like thought. *Journal of Behavior Therapy and Experimental Psychiatry*, *33*(3–4), 177–189. https://doi.org/10.1016/S0005-7916(02)00048-4

Zucker, B. G., Craske, M. G., Barrios, V., & Holguin, M. (2002). Thought action fusion: Can it be corrected? *Behaviour Research and Therapy*, *40*(6), 653–664. https://doi.org/10.1016/s0005-7967(01)00054-7