# Claims must map onto analyses:

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

# “An Empirical Investigation of Hierarchical versus Distinction Relations

# in a Self-based ACT Exercise”

# Ian Hussey

# (Abstract)

Foody et al.’s (2013) analogue study compared the efficacy of two different interventions, both based on elements of Acceptance and Commitment Therapy (ACT) (Hayes et al., 1999), in relieving distress after a distress induction procedure. Foody et al.’s (2013) study presents 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 (Hayes et al., 1999). Foody et al. (2013) 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 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. Given the importance of Foody et al.’s (2013) results to this contemporary debate, and the fact that it claims to be a replication of an original effect (thereby providing stronger evidence for the original claim), it seems important to closely examine the strength of the evidence supporting their claims.

## Summary of Foody et al. (2013)

### Stated relevance

Foody et al. (2013) state 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 timepoint: 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 VAS 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.’s (2013) hypothesis, although not stated formally, can be inferred from the study they state they were conceptually replicating (Luciano et al., 2011) and their own central conclusions. To paraphrase the purpose of their design, their hypothesis, and their claims: Foody et al. (2013) intended to induce distress (i.e., between baseline and post induction). They then wished to assess the relative degree to which the two interventions could then relieve that distress. Their hypothesis was that the “hierarchical self as context” intervention would be more effective than the “distinction self as context” intervention, and they concluded that this was the case, and 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).

## Critique

## Failure to correct for multiple comparisons

In their abstract, Foody et al. (2013) state 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 reiterate 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, only paragraphs later they state 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. Together, 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. They do not state it explicitly, but we can assume that this claim is 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 follows this pattern or has the same relevance to the claim. They report that the interaction effects were significant for stress (“*p* = .04”) but not discomfort (“*p* = .45”) or anxiety (“*p* = .33”; pp. 381-382).

Based on the magnitudes of the *p* values, it is important to note that the combination of small sample sizes, multiple outcome measures, and a subset of barely significant *p* values already represent indications that this effect and broader claim may not be replicable (REFs).

However, let us take these analyses on 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. However, 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 the most 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* = .66, anxiety: *p* = .66, stress: *p* = .12). With appropriate alpha corrections, Foody et al.’s (2013) results therefore do not support their conclusion that the hierarchy intervention more effectively relieves distress than the distinction intervention.

### Absent or inappropriate analyses

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 the scores at the baseline or post induction time points. Post hoc contrasts would be needed to explore 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 timepoints for each of the Visual Analogue items: “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 claim in their abstract and discussion that the hierarchical condition is superior to the distinction condition.

# Method

Given that Foody et al. (2013) does not report a direct and appropriate statistical inference test to support their core claim that the hierarchy condition was superior to the distinction condition, I therefore conducted a reanalysis testing this claim. Specifically, I aimed compare distress between the two groups at the post intervention timepoint using independent Welch’s *t*-tests with Hedges’ *g* standardized effect sizes (i.e., Cohen’s *d* with correction for small sample sizes). 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 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, but did not receive a reply even after follow up emails.

## Extraction of summary statistics

Independent Welch’s *t*-tests and Hedges’ g effect sizes can be 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. Although Foody et al. (2018) measured distress at baseline and post induction too, these phases of the experiment did not differ between conditions. Due to random assignment to the conditions and equivalence of conditions up until the intervention phase, any differences between the conditions can be attributed to error variance.

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 timepoint were not reported in text, only approximate values for the baseline timepoint (e.g., “<11”). However, (a) change scores for both conditions between the timepoints 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 timepoint from the plots. Means for the post intervention timepoint were calculated in two different way to validate them against one another: using the mean for that timepoint extracted from the plots; and using the mean for the baseline timepoint adding the change scores between timepoints 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, I employ 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 depends on the type of the interval. Foody et al. (2013) do 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

## Assessing 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, in order to correct for the familywise error rate, adjusted *p* values (*p*adj) were calculated using Holm corrections (see Table 2 and Figure 1).

### Analyses 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). However, the effect size for discomfort and stress were so large as to be raise questions about their credibility: both effect sizes (Hedges’ *g* = 0.99 and 0.91, respectively) are of the same magnitude as the correlation between height and weight in U.S. adults (i.e., *r* = .44, equivalent to Cohen’s *d* = XX).

### Analyses 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: discomfort (*p* = .131, *p*adj =.392), stress (*p* = .918, *p*adj =.918), and anxiety (*p* = .166, *p*adj =.392; see Table 2).

Confounded by the passage of time – no inactive placebo condition

Differences in differences

Foody et al. (2013) describe their three ad hoc VAS measures as being measures of “Distress”. 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. Putting this measurement question aside and assuming that Foody et al. (2013) are correct that their VAS measures are valid measures of distress, the mixed findings between the measures present an issue for the interpretation of the results.

Specifically, a statistically significant effect was found for the Discomfort (*p*= .005, *p*adj = .013) and Stress (*p*= .010, *p*adj = .019) outcome variables, but not Anxiety (*p*= . 826, *p*adj = . 826). 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). Whether differences in effect size between the outcome variables can be detected must be assessed directly. I did this by converting the confidence intervals effect sizes on Hedges’ g

Discomfort vs. Anxiety (*p* = .057, *p*adj = .169)

Discomfort vs. Stress (*p* = .841, *p*adj = .841)

Anxiety vs. Stress (*p* = .087, *p*adj = .173)

## Assessing differences between the conditions for a pooled outcome variable

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. It is therefore useful to also construct a single hypothesis test that

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

**Figure 1.** 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.



There are two reasons why these results are still difficult to interpret. First, the effect size for discomfort and stress are implausibly large. A wealth of research in the wake of the Replication Crisis in psychology has noted that the combination of small sample sizes and very large effect sizes, such as that observed in Foody et al. (2013), are a hallmark of unreplicable effects (REF). In providing guidance on the interpretation of both large and small effect sizes, Funder and Ozer (2019) note that a “very large effect size (*r* = .40 [Cohen’s *d* = 0.87] or greater) in the context of psychological research is likely to be a gross overestimate that will rarely be found in a large sample or in a replication.” (p. 156, effect size conversion added). The effect sizes for discomfort and stress (Hedges’ *g* = 1.00 and 0.90, equivalent to *r* = .45 and .41) are both larger than this, raising questions about them being credible estimates of the population effect. To put the effect sizes for these two outcome measures in context, this would imply that the difference in efficacy of the distinction and hierarchical conditions – both of which are plausible looking ACT-informed interventions – is roughly the same size as the correlation between weight and height in U.S. adults (*r* = .44: Funder & Ozer, 2019).

Second, even if we take the effect sizes at face value, there is an issue of interpretation.

**Table 3.** Independent Welch’s *t*-test comparing the hierarchical and distinction conditions at the post intervention time point using the pooled VAS ratings.

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

# Discussion

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.

# Key quotes

The paragraphs of results for the primary outcome measures are presented below. The key results are highlighted in bold and will be discussed further below.

### Discomfort

“Both conditions recorded similarly low levels of discomfort (<11/100) at baseline and increased to the same point after the distress induction (distinction self as context +12.7; hierarchical self as context: +16.3; see Figure 1). Thereafter, **distinction resulted in a very small increase in discomfort (+.76), while hierarchy resulted in a decrease (-7.57).** A mixed between within 3x2 ANOVA revealed a highly significant main effect for time [Wilkes Lambda = .53, F (2, 33)= 14.58, p= .000, partial eta squared= .47], but not for condition [F (1, 34)= 1, p= .32, partial eta squared= .03]. **The interaction effect was also non-significant [Wilkes Lambda= .95, [F (2, 33)= .81, p= .45, partial eta squared= .05].** Two dependent t-tests investigated which time point was influencing the significant effect. The results showed a significant increase in discomfort from baseline to post-distress induction (p= .000), but not from post-distress induction to post-intervention (p= .31).” (p. 381)

### Anxiety

“Both conditions recorded similarly low levels of anxiety (<10/100) at baseline, and both increased at post-distress induction (distinction: +8.06; hierarchy: +13.57; see Figure 2). **Anxiety subsequently decreased for both conditions, although the larger change was recorded for the hierarchical intervention (distinction: -.03; hierarchy: -3.86).** A mixed between within 3x2 ANOVA revealed a highly significant main effect for time [Wilkes Lambda= .54, F (2, 33)= 14.12, p= .000, partial eta squared= .46], but not for condition [F (1, 34)= .021, p= .89, partial eta squared= .00]. **The interaction effect was non-significant [Wilkes Lambda= .94, [F (2, 33)= .1.14, p= .33, partial eta squared= .07].** Two dependent t-tests showed a significant increase in anxiety from baseline to post-distress induction (p= .000), but not from post-distress induction to post-intervention (p= .46).” (p. 381)

### Stress

“Both conditions recorded low stress (<11/100) at baseline, and both increased at post-induction (distinction: +7.43; hierarchy: +9.48; see Figure 3). Thereafter, however, **distinction resulted in an increase in stress (+4.71), while hierarchy reduced stress (-8.82).** A mixed between within 3x2 ANOVA revealed a highly significant main effect for time [Wilkes Lambda= .65, F (2, 33)= 8.74, p= .001, partial eta squared= .35], but not for condition [F (1, 34)= .1, p= .76, partial eta squared= .00]. However, **the interaction was significant [Wilkes Lambda= .83, [F (2, 33)= 3.46, p= .04, partial eta squared= .17].** Two dependent t-tests showed a significant increase in stress from baseline to post-distress induction for both conditions (both ps <.05). The increase in distinction from post-distress induction to post-intervention was not significant (p= .23), but the decrease for hierarchy at the same time point was significant (p= .02).” (p. 382)