Robustly null:
A critical reanalysis of Foody et al. (2013) and Foody et al. (2015)

Ian Hussey

The evidential link between Relational Frame Theory (RFT) and Acceptance and Commitment Therapy (ACT) is a matter of on-going debate. Foody et al. (2013) is frequently cited as evidence for ACT's concept of Self-as-Context and as evidence for the ties between ACT's middle level clinical terms and the more basic account of language and cognition provided by RFT. However, the failed replication study by Foody et al. (2015) typically goes uncited. Given their importance, I critically re-evaluate the results of both articles. I found several issues, the most important of which was that the central claim does not correspond with the reported analyses. When I extracted summary statistics from both studies and used them to conducte a multiverse analysis, that mapped directly onto the original claim (i.e., that the hierarchy condition reduced distress more than the distinction condition), null results were found under every set of conditions. Taken together, results illustrate how weak or flawed research practices and uncritical citation of results can hinder cumulative science and evidence based clinical practice.

A long-standing debate within the literature on Acceptance and Commitment Therapy (REF) is whether or not ACT's model of therapeutic change is or is not linked to the more basic science of learning and behavioral processes, specifically relational responding, described by Relational Frame Theory (REF). Foody et al. (2013) argued that they were among the first to successfully attempt to bridge the gap between ACT's model of therapeutic change 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). 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).

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, however, 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. As such, there appears to be 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).

The impact of Foody et al. (2013, 2015)

Foody et al. (2013) represents a key article 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). At the time of writing it has 122 citations on Google Scholar.

Foody et al. (2013) is often cited as evidence for (a) the superiority of hierarchy-based interventions over distinction-based interventions, and therefore (b)

as evidence for ACT's concept of Self-As-Context. More generally, Foody et al. (2013) is often cited as evidence of (c) the link between RFT and ACT, in general. I provide several examples below. These examples are intended to be illustrative rather than a systematic review. I have selected them from sources that are likely to be particularly influential such as textbooks, articles providing introductions to the topic, and articles advocating for the expansion of ACT training (e.g., based on its putative links to RFT). In these examples, it is worth noting the pattern that the articles are cited without interrogating the strength of the evidence actually provided by Foody et al. (2013) and without acknowledging that these results failed to replicate in Foody et al. (2015).

Foody et al. (2013; 2015) are cited in three different chapters of the recently published Oxford Handbook of Acceptance and Commitment Therapy (REF). In their chapter on "Cognitive Defusion", a key concept within ACT, Ruiz et al. (2021) state *"Foody, Barnes-Holmes, Barnes-Holmes, and Luciano (2013) found that Defusion II [hierarchy] was more efficacious in reducing experimentally induced emotional distress than Defusion I [distinction]."* This takes the conclusions of Foody et al. (2013) at face value without interrogating the strength of the evidence it provides for the claim. Additionally, they erroneously state *"Foody, Barnes-Holmes, Barnes-Holmes, Rai, and Luciano (2015) found that protocols that included framing ongoing private events through hierarchical relations were more efficacious than those that only introduced deictic relations."* (p. 13). This mischaracterizes the null results found by Foody et al. (2015) as if they support the claim.

In their chapter on "Clinical behavior analysis and RFT: Conceptualizing psychopathology and its treatment", Luciano et al. (2022) cite Foody et al. (2013) as supportive evidence in the section on *"Evidence of hierarchically framing ongoing behavior as a central relational process"*. Again, this takes the conclusions of Foody et al. (2013) at face value without interrogating the strength of the evidence it actually provides for the claim. Additionally, they do not cite the failed replication by Foody et al. (2015).

In their chapter on "A primer on Relational Frame Theory", Harte & Barnes-Holmes (2022) cite Foody et al. (2013) as a demonstration of hierarchical framing, again without interrogating the strength of the evidence it provides, and do not cite the failed replication by Foody et al. (2015). It is important to recognize that there is significant overlap in authorship between these publications, so mere unawareness of the failed replication is implausible.

Foody and colleagues' work is also cited in other influential textbooks. In "Behavior Therapy" (REF), the chapter "The Future of Third Wave Cognitive Behavior Therapies" (Zettle & Masuda, 2022) cites both Foody et al. (2013) and Foody et al. (2015) as an illustrative example of the link between RFT and ACT. However, they do not either interrogate the strength of the evidence or mention that the latter article is a failed replication. Additionally, they cite Sierra et al. (2016) without citing or discussing the failed replication of that study by Pendrous et al. (2020).

Many other examples can be found. In their article advocating for the inclusion of ACT training in Applied Behavior Analysis curricula, Kelly & Kelly (2021) cite Foody et al. (2013) as evidence of the link between RFT and ACT. They do not examine the strength of the evidence provided by Foody et al. (2013) or cite the failed replication by Foody et al. (2015). In their article on the same topic, Dixon and Hayes (2022) cite both articles as supportive evidence without acknowledging that the latter is a failed replication. In their article celebrating the contributions of Murray Sidman, Law & Hayes (2021) state *"understanding how to foster healthy perspective-taking seems central to establishing self-direction, independence, and values-based actions (Foody et al., 2015)"*, which not only mischaracterizes the nature of this failed replication but also the general nature of the study and its possible conclusions, which were not related to self-direction, independence, or values-based action. Unfortunately, recent research has demonstrated that such inaccurate and biased citations are surprisingly common in the psychology literature, with roughly 9% of all citations checked grossly misrepresenting the original work (REF).

Even replication studies, which might be expected to be more aware of the importance of citing failed replications, fall into this pattern. Gomide et al. (2024) cite Foody et al. (2013) favorably, without interrogating the strength of the evidence it provided and without citing the failed replication by Foody et al. (2015).

Of course, it is possible to find examples of more critical readings of Foody et al. (2013). Often, these come from researchers and journals that are not already strongly aligned with ACT and RFT. For example, Godbee & Kangas (2022) are more circumspect, stating *"Although the differences between hierarchical [Self-as-Context] ("I am more than my experiences") and distinction [Self-as-Context] ("I am not my experiences") have been researched, there is limited evidence that one type is more effective than the other (Atkins & Styles, 2016; Foody et al., 2013; Foody et al., 2015)."*

Is such a critical reading of the results of Foody et al. (2013) necessary? In the next section, I briefly summarize Foody et al.'s (2013) purpose, design, results and conclusions. I then provided critiques of the evidence it presents. Lastly, I present a re-analysis of the results of both Foody et al. (2013) and Foody et al. (2015) to assess whether these studies can provide robust evidence for their claims.

Commented [IH1]: Some book chapters are 2021 and some are 2022; correct these to the book year

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: baseline, post distress induction, post ACT intervention) X 2 (between: "hierarchical self as context" vs. "distinction self as context" intervention) 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). Other measures (i.e., the Acceptance and Action Questionnaire II, delivered only at baseline, and a Reactions Questionnaire that served as a manipulation check) are not considered here. Participants completed the VAS at baseline, then completed a distress induction task, then were assessed again (post distress induction), then completed an ACT intervention (randomized 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 intervention 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). The key statistical results they provide to support this claim are the interaction effects between time point and group on the 3 X 2 RM-ANOVAs. A statistically significant result was found for one of the three outcome measures (stress, $p = .04$; anxiety, $p = .45$, discomfort, $p = .33$). No statistical tests compared the groups at specific time points.

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 hierarchical intervention only resulted in a reduction in all three dependent measures, including a significant reduction in stress."* (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). I understand this last quote as stating that all three outcomes showed descriptive differences in the expected direction, although only stress was statistically significant.

Critique of the original method & analyses

Status as a replication

Foody et al. (2013) state that "the research was a replication of a previous study by Luciano et al. (2011)". The appeal to being a replication study can lend additional credibility to claims. There is no single universally accepted definition of a replication, and there is ongoing debate about what nomenclature and taxonomy is useful in distinguishing between subtypes such as conceptual versus direct replications (e.g., REFs). Typically, I do not find this definitional debate to be particularly useful. However, in this case, I think it is worth pointing out how little overlap there is between Luciano et al. (2011) and Foody et al. (2013), and that by all reasonable accounts Foody et al. (2013) cannot credibly or usefully be described as a replication, in the sense of the weight of evidence it presents by being labelled as such. Of course, Foody et al. (2013) do acknowledge one important difference between their work and Luciano et al. (2011) when self-identifying the work as a replication: *"the research was a replication … except that we were able to use less intensive interventions with our non-clinical sample"* (p. 384). However, many other differences between the study exist but were not acknowledged in Foody et al. (2013), including the sample, design, number of follow up period, outcome measures, and analytic strategy (see Table 1).

Table 1. Comparisons between Luciano et al. (2011) and Foody et al. (2013)

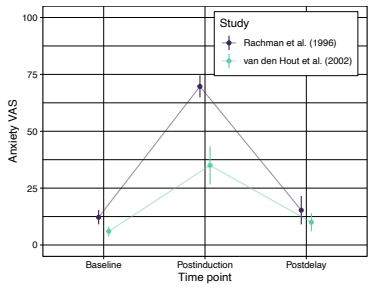
	Luciano et al. (2011)	**Foody et al. (2013)**
Sample	Adolescents	Adults
Design	Quasi-experiment (non-random assignment to defusion intervention conditions based on severity)	Laboratory analogue study (random assignment to defusion intervention condition)
Distress induction	No	Yes
Interventions	Value clarification (all participants) Defusion (distinction vs. hierarchical conditions; non-random assignment)	Defusion (distinction vs. hierarchical conditions)
Timepoints	5 therapy sessions over 4 months	Single experimental session
Follow up period	2 weeks and 4 months	Single experimental session
Outcome variables	Spanish Avoidance and Fusion Questionnaire (AFQ-S) Accepting without Judgment Scale of the Kentucky Inventory of Mindfulness Skills (KIMS) Impulsive Behavior Inventory (IBI)* Emotional Behavior Inventory (EBI)*	Single-item Visual Analogue Scales for discomfort, anxiety, and stress*
Primary analysis	Mann-Whitney *U* tests	RM-ANOVAs fitted to all three time points between conditions

Notes: *Ad hoc measures created for the study.

Inconsistent descriptions of the methods

Both Foody et al. (2013) and Foody et al. (2015) report results from a Reactions Questionnaire that includes references to a car accident, e.g., "Please rate how vivid your thoughts and images were of the car accident" (REFs). However, neither study's method section includes any reference to a car accident. The distress induction procedure in both is stated to be requiring the participant to write out a negative self-referential thought. One possibility is that text from a different study was erroneously copied into the wrong manuscripts. For example, a previous publication by Foody et al. (2012) compared two versions of a different distress induction procedure that involved having the participant write the sentence "I hope [participants' loved one] is in a car accident" (REF). It is unfortunate that these unexplained references to car accidents got through peer review at two different journals apparently without being detected or corrected.

Figure 1. Results from no-intervention negative control conditions reported in previous distress induction studies. Points represent means, error bars represent 95% Confidence Intervals.



No control condition

Both Foody et al. (2013) and Foody et al. (2015) are seriously undermined by the lack of a control condition that did not receive an intervention. Control conditions are of course needed to determine casualty within an experimental approach. Even more than usual, however, the need for a control condition becomes especially obvious when studying effects that are due to short term manipulations of mood, as in Foody et al.'s work. Mood inductions tend to produce very short-lasting effects, and mood tends to return to baseline levels within a very short period of time. As such, in order to be able to differentiate between any change in mood due to the intervention vs. due to natural improvement over time, a negative control condition is needed (i.e., where participants are given no intervention).

This is not case of "hindsight is 20/20" or a post hoc call for rigor: the authors of Foody et al. (2013, 2015) developed their distress induction procedure from work that demonstrated this natural return to baseline levels over a short period of time (e.g., 2 minutes). In an article they published the previous year which compared versions of distress induction procedures, Foody et al. (2012) cited work by Rachman et al. (1996) and van den Hout et al. (2002). Like both Foody et al. (2013) and Foody et al. (2015), these studies recorded anxiety scores using a Visual Analogue Scale, and measured affect at baseline and after a distress induction procedure. However, they also included no-intervention conditions to study the natural recovery of baseline mood. Both studies demonstrated that mood returned to baseline after a short delay (two minutes) in the absence of any intervention (Rachman et al., 1996: Hedges' g_s = 3.24, 95% CI [2.51, 3.97]; van den Hout et al., 2002: g_s = 1.17, 95% CI [0.69, 1.64]). I extracted the summary statistics from Rachman et al. (1996) and van den Hout et al. (2002) and present them

in Figure 1. Note the pattern of effect across timepoints can be seen: a baseline participants demonstrate low anxiety, at postinduction it is higher, and at post-delay it returns nearly to baseline. However, these two studies involved no intervention: the reduction in distress after the delay was spontaneous. Foody et al. (2013) did not include any such control condition where no intervention was provided, and did not discuss the highly plausible possibility that any reductions in distress after intervention were merely due to spontaneous recovery rather than due to their intervention. At the same time, it is plausible that Foody and colleagues were aware of this possibility given that they cited both Rachman et al. (1996) and van den Hout et al. (2002) in their earlier publication (Foody et al., 2012). As such, in the absence of a control condition, it would be erroneous to conclude from Foody et al.'s (2013, 2015) results that the interventions caused reductions in distress between timepoints.

How to interpret the results

Original results represent weak support

We can reasonably 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 (i.e., of one significant result and two non-significant results between outcomes) or has the same degree of 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).

[Add results of Foody et al 2015 here]

Applying multiple testing corrections produces null results

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). 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, significant results were obtained for only one of the three outcome measures, without the use of any familywise error corrections. Foody et al. (2013) therefore make a disjoint claim, where the alternative hypothesis is accepted on the basis of one or more positive results among multiple tests (whereas a joint claim would require positive results on every test).

For individual tests, the false positive rate is equal to the alpha value (e.g., 5%) when the test's assumptions are met. However, disjoint claims suffer from increased false positive rates, because even a single false positive will cause an incorrect claim. With three outcome measures and an alpha value of 5%, the familywise false positive rate for the general claim can be as high as 14.3% (i.e., when tests are fully independent).

$$FPR = 1 - (1 - \alpha)^k \quad (1)$$

When disjoint claims are made, familywise error corrections are required to maintain nominal false positive rates implied by the tests' alpha level. If one is willing to accept the alternative hypothesis on the basis of any significant result across multiple outcome measures (i.e., a disjoint claim), each of which has a risk of false positives, some form of familywise alpha correction is necessary to keep the long run false positive rate within the nominal alpha value (e.g., 5%). Foody et al. (2013) do not employ any alpha corrections, but they can be applied post hoc to the reported p values.

Applying even a relatively 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.

[Add results of Foody et al. 2015 here]

The RM-ANOVAs are not suitable to testing the original claim

Regardless of the results obtained (i.e., even if all results had been statistically significant), I argue that the RM-ANOVAs cannot by themselves support the claims made about the superiority of the hierarchy condition. By including the baseline scores, the interaction effects do not test the hypothesis that the interventions produce differential effects, because interaction effects could be driven by one or more of the baseline, post induction or post intervention time points. That is, a very different pattern of results that does not represent the superiority of the hierarchy condition in reducing experimentally induced distress could nonetheless produce significant p values for the interaction effect.

In order to support their claim, an elaborated or alternative analytic strategy would be needed. For example, post hoc contrasts could have been used to test for differences in means between the groups at each time points; or to differences in means at the post-intervention time point could have simply been assessed via a t-test; or an ANCOVA could have been used to compare differences in means at the post-intervention time point while controlling for differences at baseline. However, Foody et al. (2013) do not report any such results. 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.

Reanalysis

In this section, I report the results of a reanalysis that sought to provide a direct test of Foody et al.'s (2013) central aim, i.e., *"The primary aim of the current study was to compare the relative utility of the two self-based interventions (distinction versus hierarchical relations) in reducing participants' discomfort, anxiety, and stress after exposure to the distress induction task."* (p. 381). Specifically, I aimed to 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).

Of course, data analysis involves multiple decisions, each of which may have more than one plausible and defendable choice. In order to understand the robustness of conclusions across multiple plausible choices, the concept of "multiverse analyses" (sometimes called "specification curves"; Orben & Przybylski REF). I will return to this analysis after first explaining how results were extracted from the articles.

Transparency and data availability

All raw and processed data as well as R code for data processing and analyses are available (osf.io/ztd8n).

Attempts to obtain the original data

In the first instance, I attempted to obtain the original data. I contacted all authors of Foody et al. (2013) requesting the data. Unfortunately, the first author informed me that the dataset no longer exists. Luckily, however, such tests could be conducted based on the summary statistics presented in the original article.

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. Results for both the distinction and hierarchy conditions were extracted from Foody et al. (2013). I also extracted results for the same two conditions that were replicated in Foody et al. (2015). The two additional conditions included in Foody et al. (2015: i.e., object-distinction and object-hierarchy) were not extracted, as I focused only on the effects that were replicated. Similarly, although Foody et al. (2015) employed a second cycle of distress induction and intervention, I examined only the first phase of distress induction and intervention that replicated what was done in Foody et al. (2013).

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)." (Foody et al., 1013, p. 375); "[quote]" (Foody et al., 2015, p. XX).

For the original study reported in Foody et al. (2013), 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. Means for the post intervention time point were therefore calculated in two different way to validate them against one another: using the mean for that time point extracted from the plots using WebPlotDigitizer (Marin et al., 2017); 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 the estimates obtained via the XXX method for the reanalyses.

Standard Deviations for Foody et al. (2013) were recalculated from the Standard Error intervals reported in their Figures 1 to 3 and their sample sizes using the below equation. Data extraction was much simpler for Foody et al. (2015) as all means and SDs were reported in Table 2.

$$SD_{SEM} = \frac{interval\ width}{2} \times \sqrt{\bar{n}} \qquad (2)$$

In order to compare differences between the distinction and hierarchy conditions at the post-intervention time point, while also acknowledging that there may be legitimate analytic

Hypothesis tests 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 Errors of the differences in means (SE), t values (t), and degrees of freedom (df) using the below equations.

$$SE = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}} \qquad (3)$$

$$t = \frac{M_1 - M_2}{SE}$$

$$df = \frac{\left(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}\right)^2}{\frac{SD_1^2}{n_1 - 1} + \frac{SD_2^2}{n_2 - 1}} \qquad (4)$$

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.

$$g = \frac{M_1 - M_2}{\sqrt{\frac{SD_1^2 + SD_2^2}{2}}} \times \left(1 - \left(\frac{3}{4 \times (n_1 + n_2 - 2) - 1}\right)\right) \quad (5)$$

p values and multiple testing corrections

When performing multiple tests of a more general hypothesis using multiple correlated outcome measures, as here where measures of anxiety, discomfort, and stress are used to make conclusions about distress more generally, researchers are typically advised to maintain the severity of their hypothesis test (i.e., their false positive rate) using familywise error corrections. Although not doing so provides a weaker test, this may still represent a legitimate choice between researchers (i.e., some may be more willing to provide a weaker test). I therefore calculated both p values (using t and df) and Bonferroni-corrected p values. For Hedges' g effect sizes, I calculated both 95% Confidence Intervals and 98.33% Confidence Intervals, which correspond to the correct Bonferroni adjustment for three outcome variables (i.e., $1 - \lfloor 0.05 \ / \ 3 \rfloor$). Both intervals were employed in the multiverse analysis to assess the robustness of conclusions to this analytic choice.

Hypothesis tests using a pooled outcome measure of distress

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 with regards to their three component outcome measures (discomfort, anxiety, and stress). Whether or not it is appropriate from a measurement perspective to treat these three ad hoc measures as valid measures of a latent "distress" variable cannot be answered without access to the original data (or possibly via new data collection). Nonetheless, it is possible to calculate a single pooled outcome measure to test this more general claim. This would also serve to avoid any potential issues with regard to how to interpret a mix of significant and non-significant results between outcome measures, as observed in Foody et al. (2013), where significant results were found for only one of three outcome measures ("stress") but the authors made more conclusions about "distress". I therefore I calculated pooled means for each condition by averaging them, and pooled Standard Deviations using the following formula. These were used to calculate a further set of Welch's t-tests and effect sizes for both the original study and the replication.

$$SD_{pooled} = \sqrt{\frac{SD_1^2 + SD_2^2 + SD_3^2}{3}} \quad (6)$$

Note that because the pooled effect size provides a single test of the hypothesis, no Bonferroni-adjusted version of the Confidence Intervals was corrected. All four outcome measures (anxiety, discomfort, stress, and the pooled outcome measure) were employed in the multiverse analysis in order to attempt to understand the robustness of conclusions to the choice of outcome measure.

Controlling for baseline

A common analytic decision that must be made when analyzing differences between conditions after an intervention is whether or not to the control for differences at baseline (REF). Given only access to the summary statistics and not the original data (i.e., in order to estimate the correlation between time points), it was nonetheless possible to use Morris corrections (REF) to control for these differences (i.e., corrected Hedges' $g_{corrected}$ = Hedges' $g_{postintervention}$ – Hedges' $g_{baseline}$). Both Hedges g (not correcting for baseline scores) and Hedges $g_{corrected}$ (corrected for baseline scores) were entered into the multiverse analysis in order to assess the robustness of conclusions to this analytic choice.

Results

Multiverse analyses involve calculating a metric under all permutations of analytic choices using data from a given underlying dataset. In this case, the outcome metric was Hedges g (including its variation controlling for baseline scores), and the analytic choices were 1) the outcome measure (anxiety, discomfort, stress, and pooled), 2) whether baseline scores were controlled for or not, and 3) whether corrections for familywise error rate were used to choose the width of the Confidence Interval (95% vs. 98.33%, corresponding to a Bonferroni correction for the three correlated outcomes).

Figure 2 presents the results of the first multiverse analysis for the data extracted from original study by Foody et al. (2013). The upper panel presents the distribution of Hedges g effect sizes arranged from lowest to highest values. The lower panel presents the analytic choices that gave rise to each effect size (i.e., the one presented directly above it). Figure 3 presents the results of the same multiverse analysis applied to data extracted from the replication study by Foody et al. (2015). As can be seen from the figures, specifically the failure of all Confidence Intervals to exclude the zero point, no significant differences were found between the distinction and hierarchy conditions at the post-intervention time point under any set of analytic choice in either the original study or the replication. That is, the results very robustly fail to support the original conclusion that *"The findings demonstrated superiority of the intervention that focused on hierarchical, rather than distinction, deictic relations in terms of reducing distress."* (p. 373). This result is in strong contrast with the conclusions of Foody et al. (2013).

Figure 2. Multiverse plot for the original study (Foody et al., 2013).

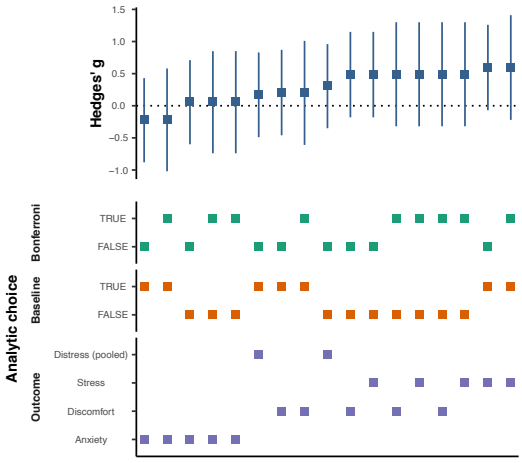


Figure 3. Multiverse plot for the replication study (Foody et al., 2015).

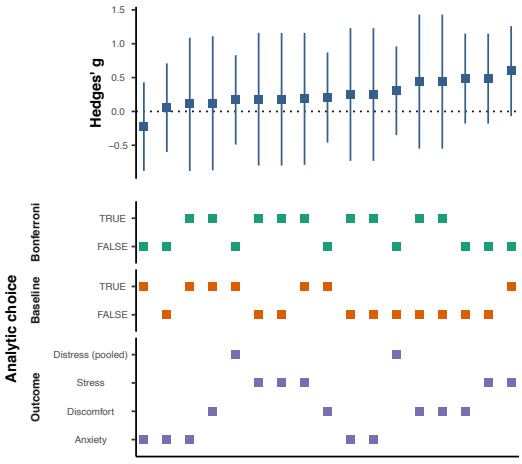
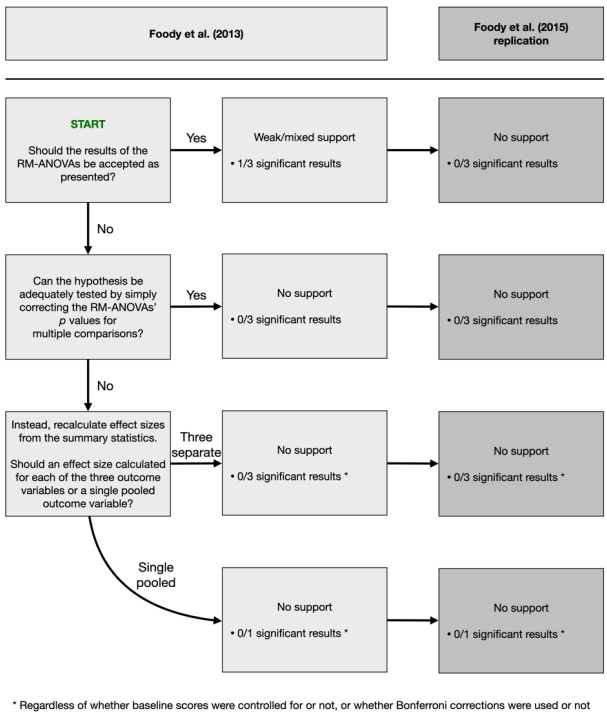


Figure 4. Flow chart for the interpretation of results from Foody et al. (2013) and Foody et al. (2015)



Discussion

The combined results of the critiques and reanalyses are presented as a decision chart in Figure 4. Starting from the top left, the reader can make one analytic decision at a time: starting with whether the results of Foody et al.'s (2013) RM-ANOVAs should simply be accepted on face value; if not, followed by whether merely correcting their results for multiple testing is sufficient; and if not, followed by which form of between groups effect size should be calculated. In summary, no set of analytic choices for Foody et al. (2013) followed by applying the same choice to the replication study in Foody et al. (2015) produce (a) a replicable result, and (b) all choices other than accepting Foody et al.'s (2013) RM-ANOVA results provide null results for all tests in both studies. Even for that exception, support for the hypothesis is limited to one out of three outcome measures, undermining Foody et al.'s (2013) general claim. This is remarkable given the way that the results of Foody et al. (2013) are referred to in other work – i.e., as evidence for the utility of RFT to ACT, and support for the ACT model of psychotherapy.

The literature is (currently) rendering it as truth that hierarchy is superior to distinction, and that therefore ACT's model of self-as-context is true due in part to Foody et al. 2013, and that therefore ACT benefits from its close links to basic RFT research. Closer reinspection suggests that the first link in this evidential chain is not the case, which should therefore undermine all subsequent links in the belief chain. However, correcting these erroneous beliefs in clinicians and scientists will likely be far from trivial.

Not an isolated case

The collective rendering by the CBS community of Foody et al. (2013)'s null findings into an incorrect but widely believed scientific truth should give us pause for thought. Some may be quick to suggest that this is a single unfortunate example. It may be worth recalling that similar appeals to exceptionalism were made early in the replication crisis, and yet the problems kept spreading to new areas of work people previously argued to be unaffected. We must be careful to avoid hubris here too, and instead be open to the possibility that many areas of our work are built on extremely shaky foundations.

Worryingly, 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).

Commented [IH2]: Early draft content

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/

Dixon, M. R., & Hayes, S. C. (2022). On the Disruptive Effects of Behavior Analysis on Behavior Analysis: The High Cost of Keeping Out Acceptance and Commitment Therapy and Training. *Behavior Analysis in Practice*, 1–7. https://doi.org/10.1007/s40617-022-00742-4

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/

Foody, M., Barnes-Holmes, Y., Barnes-Holmes, D., & Luciano, C. (2013). An Empirical Investigation of Hierarchical versus Distinction Relations in a Self-based ACT Exercise. *International Journal of Psychology*.

Foody, M., Barnes-Holmes, Y., Barnes-Holmes, D., Rai, L., & Luciano, C. (2015). An Empirical Investigation of the Role of Self, Hierarchy, and Distinction in a Common Act Exercise. *The Psychological Record*, *65*(2), Article 2. https://doi.org/10.1007/s40732-014-0103-2

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/

Godbee, M., & Kangas, M. (2022). Focusing on the self in context as an emotion regulatory strategy: An evaluation of the "self-as-context" component of ACT compared to cognitive reappraisal in managing stress. *Anxiety, Stress, & Coping*, *35*(5), 557–573. https://doi.org/10.1080/10615806.2021.1985472

Gomide, C. P., Perez, W. F., & Pessôa, C. V. B. B. (2024). Perspective taking reduces the correspondence bias: A systematically replication of Hooper et al. (2015). *Journal of Contextual Behavioral Science*, *32*, 100735. https://doi.org/10.1016/j.jcbs.2024.100735

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

Harte, C., & Barnes-Holmes, D. (2022). A primer on relational frame theory (RFT). In M. P. Twohig, M. E. Levin, & J. M. Petersen (Eds.), *The Oxford Handbook of Acceptance and Commitment Therapy* (1st ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197550076.001.0001

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

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

Kelly, A. D., & Kelly, M. E. (2021). Acceptance and Commitment Training in Applied Behavior Analysis: Where Have You Been All My Life? *Behavior Analysis in Practice*, *15*(1), Article 1. https://doi.org/10.1007/s40617-021-00587-3

Law, S., & Hayes, S. C. (2021). Murray Sidman: Fostering progress through foundational choices. *Journal of the Experimental Analysis of Behavior*, *115*(1), 21–30. https://doi.org/10.1002/jeab.640

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*.

Luciano, C., Törneke, N., & Ruiz, F. J. (2022). Clinical Behavior Analysis and RFT: Conceptualizing Psychopathology and Its Treatment. In M. P. Twohig, M. E. Levin, & J. M. Petersen (Eds.), *The Oxford Handbook of Acceptance and Commitment Therapy* (1st ed., pp. 109–142). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197550076.013.5

Marin, F., Rohatgi, A., & Charlot, S. (2017). *WebPlotDigitizer, a polyvalent and free software to extract spectra from old astronomical publications: Application to ultraviolet spectropolarimetry* (arXiv:1708.02025). arXiv. http://arxiv.org/abs/1708.02025

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., Gil-Luciano, B., & Segura-Vargas, M. A. (2021). Cognitive defusion. In M. P. Twohig, M. E. Levin, & J. M. Petersen (Eds.), *The Oxford Handbook of Acceptance and Commitment Therapy*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197550076.001.0001

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

Zettle, R. D., & Masuda, A. (2022). The Future of Third Wave Cognitive Behavior Therapies. In W. O'Donohue & A. Masdua (Eds.), *Behavior Therapy*. https://doi.org/10.1007/978-3-031-11677-3_34