**Reviewer 1 Feedback:**

This reviewer provided high-level feedback regarding our manuscript, which makes it difficult to respond to point-by-point as I had with other reviewers. As such, we will go through the review line-by-line to address each critique but will note if the point had already been addressed in a previous response. All following references to page numbers will refer to the tracked changes version of the manuscript:

“*In Study 1, the authors discuss their hypotheses related to 'intensity' as if intensity was an independent variable that's being manipulated (as it is in Gal Sheppes foundational work) but this may not be appropriate in Study 1. In this study, intensity is a measured variable (not a manipulated one)…”*

**Our response:** This critique, as we have read it, argues that we have inappropriately framed our predictor variable of interest as experimentally manipulated rather than observed. However, we are not sure whether the reviewer interpreted our variable in this way because of the language we had used or because of our comparison to Sheppes’s work. As such, we respond to both.

We make comparisons to “Sheppes[’s] foundational work” not because we believe that standardized ratings represent an apples-to-apples comparison to idiographic self-report metrics, but simply because it is the preeminent standard to which almost all emotion regulation strategy research is currently being contrasted to. With very few other studies, as recently noted by Specker, Sheppes, and Nickerson (2024), in the extant literature that use subjects’ self-reported emotional experiences as predictors of emotion regulation, the Sheppes paradigms remains the closest conceptual contrast to contextualize our results, even if we deviate from one another regarding experimental control. While having access to standardized ratings would have been an excellent supplementary analysis (and we discuss reasons why we weren’t able to achieve in that in edits to page 53), it alone would not be in line with the larger goals of this project. As we view it, the models are still similar in the phenomena and associations they intend to represent – using emotional intensity (whether that be standardized or non-standardized) to predict emotion regulation (whether that be usage or choice); they primarily differ in how they prioritize external and internal validity when modeling the target phenomena. Therefore, we think that the comparison is valid, important, and does not on its own give the impression that our predictor in Study 1 was experimentally manipulated. However, to make this as clear as possible, we added additional statements to emphasize the differences between this study and previous studies on pages 6, 9, 20, 27, 31, and 53.

In response to this critique, we had also reviewed the language we used to describe this variable. We did identify any language that, to us, seemed to imply experimental control, but we would be happy to make corrections if specific examples were cited. We do explicitly note in our manuscript that our approach is observational and that Study 1 emotion intensity is not experimentally manipulated (e.g., “*It must be noted that we did not directly manipulate emotional intensity within this design*…”, pg. 53). When describing both our hypotheses and model compositions, we are careful to avoid language that suggests causality or experimental control. For example, we use terms like ‘associated’ (“*Study 1 tested whether the emotional intensity of negatively-valenced events* ***was******associated with*** *the likelihood …*”*,* pg. 11) rather than terms like ‘causes’. Lack of experimental control was a concern cited by previous reviewers (See our exchanges with Reviewers 3) and we had previously modified our manuscript to most accurately reflect our observational approach, but there may still be some examples of inaccurate language, which again if noted, we would be happy to correct.

*“…and in this particular study context, it's very likely an outcome of regulation just as much as it's a predictor of regulation.*”

**Our response:** This is an important point and was not previously sufficiently discussed in our manuscript. We agree that we are unable to disentangle the extent to which our measured variable is a product of or precursor to emotion regulation in this design. While we obviously intended to target emotion as a precursor and chose language ideally trying to target it as such, there are reasons why post-regulation affective assessments may have been reported instead or may have biased reporting. We have updated our text on pages 28 and 53 to reflect this limitation. We have also added the precise language used to capture self-report in the interest of transparency (pg. 17).

To attempt to address the reviewer’s concerns regarding experimental control and limitations of our primary predictor, we conducted an additional analysis examining regulation usage as predicted by the section of the haunted house in which the regulation occurred – a proxy for emotion intensity. The organizers of the haunted house designed two sections to be high-intensity and two sections to be low-intensity (See Cliver et al., 2024 for more details). These constitute a more controlled, albeit a lower resolution, representation of emotion intensity than our primary analysis. An assessment of self-reported fear collected immediately after each section mirrored this design structure. Using these sections as categorical predictors allowed us to explore associations with strategy usage in a similar fashion to the Sheppes and Sheppes-inspired work, but still resulted in no observed association between intensity and usage. This is of course an imperfect, post-hoc solution to this criticism, especially given that the majority of regulated events occurred in high-intensity sections, but we believe it at least adds credence to the notion that our null results in Study 1 are not simply a product of lack of control. These changes are covered on pages 19 and 27.

*“For example, another way to interpret figure 4 is that - rather than this being an unexpected finding - it might be exactly what one would predict if emotional intensity was the \*outcome\* of regulation success, which it very well could be given that these variables were measured at approximately the same time: as regulation success decreased, emotional intensity increased.”*

**Our response:** I am not sure that I follow the logic that emotion intensity measured as a product of self-regulation would alter the relationship between intensity and regulation success as we observed. As reported by the previously cited Specker paper and supporting literature within that manuscript, pre- and post-regulation intensity reports are highly correlated (r = 0.809, p < 0.001). Additionally, the difference in intensity of high and low intensity events are still significant even when measured at post-regulation (t(127) = -25.9, p <0.001). This leads me to believe that whether measured as a pre- or post-regulation assessment, distraction should be associated with events of a greater intensity by our most liberal estimations or of no intensity difference at our most conservative estimations.

If we were modeling the canonical association between success, intensity, and strategy, I would expect to see an interaction, in which reappraisal demonstrates a negative association between intensity and success and distraction demonstrates a positive association, but we do not see evidence for that. At all intensity levels that we have observations for, reappraisal appears to be more successful than distraction in this context. As stated in the manuscript (pg. 29), we have no evidence to suggest that this generalizes beyond the highly specific features of this context and this evidence in itself does not conclusively explain what we observed within this context (pg. 30). As such, we have kept our initial interpretation of this result within the manuscript but have also incorporated the reviewer’s interpretation.

As noted in recent edits (See pg. 20), both a controlled and observational approach demand different assumptions and we made design decisions in pursuit of measuring this effect while attempting to maximize ecological validity. This obviously comes with many confounds, but is an important litmus test: if a seemingly well replicated and highly robust effect can’t be observed in ways and in conditions that an everyday lay person might interpret it, what is its pragmatic value? The results of many highly controlled lab studies speak as if standardized ratings represent idiographic lived experience, as if I should use distraction in situation A and reappraisal in situation B, but when used in that way, we fail to find congruence. This study exists to demonstrate the practical boundaries of highly controlled experimental research on this topic.

We agree that intensity captured as a standardized rating (as occurs in the Sheppes paradigms) and intensity captured as an idiographic self-report metric (as occurred here) is an apples-to-oranges comparison. However, both are imperfect representations of the same thing: experienced affective intensity. We only make comparisons between the two insofar as this. If psychology

, but one that has to occur if psychology is to pursue pragmatism. Both represent some facet of experienced affective intensity, or at least that’s how I believe an audience interprets them.

This is all suppositional at this point, so this serves as a prior which others can build off of.

which we believe constitute a valuable contribution to the field, given the lack of extant literature attempting to translate this already well-trodden effect beyond

*“For this reason, I wouldn't necessarily feel comfortable with this take-home message, which makes it sound like the present results are the opposite of what prior work has demonstrated: "Though the extant literature from comparable lab studies should motivate us to expect the efficacy of distraction to increase and reappraisal to decrease as affective intensity increases, our data seems to document a deviation from this pattern in a high-intensity, quasi-naturalistic setting: distraction appeared to be less - not more - successful as affective intensity increased.”*

**Our response:**

*“Aside from this moderation effect, the main effect between intensity and strategy choice was the primary analysis in this study, and there was no reliable/significant association found. That could be informative but, given the nature of the study, I'm not sure how informative this null association is. If emotional intensity drives use of distraction vs. reappraisal, we'd expect a positive association between intensity and distraction use. But if [lower] emotional intensity is also the \*outcome\* of successful regulation - especially distraction, which was the modal strategy used - we'd expect a negative association between intensity and distraction use. These two patterns operating at once could yield a null result, which could explain Study 1 findings.”*

**Our response:** We strongly believe that a lack of analogous studies does not represent a lack of value in a non-standardized approach. Rather, a non-standardized, less-controlled approach should be employed to assess the pragmatic value of statistically robust and well-established effects if psychology is to have practical utility.

*“This alternative interpretation also tracks with what the authors found in Study 2: when new participants are told that a given event is higher vs. lower intensity (i.e., intensity is manipulated here, rather than measured like in Study 1), they choose distraction (vs. reappraisal) more often.”*

**Our response:**

*“This is essentially a conceptual replication of the Sheppes work because intensity is manipulated (i.e., given to participants) and isn't really comparable to the intensity variable in study 1, which is a complex experience that is likely being affected by regulation as much as it's affecting regulation.”*

**Our response:**

*“For this reason, Study 2 can't effectively be used to help explain the pattern of results from Study 1.”*

**Our response:** We completely agree, which is why we stated.

*“The authors then conducted Study 3 to learn whether the link between intensity and distraction choice (vs. reappraisal choice) would be present in forecasted regulation contexts (like Study 2) but not in executed regulation contexts (like Study 1).”*

**Our response:** You misunderstand.

*“But by my read, this isn't the core difference in the findings between Study 1 and 2 and so when I saw that the experimenters were again manipulating intensity in Study 3 (this time with pre-piloted lower vs. higher intensity film clips), it seemed fully reasonable for them to replicate the 'canonical relationship' between intensity and distraction, which they did. This pattern makes good sense if Study 2 and 3 are interpreted as solid conceptual replications of the original Sheppes work, where intensity is carefully manipulated for participants.”*

**Our response:**

Based upon the highlighted passages, we interpreted this feedback as containing three primary but related points, which we will response to here:

1. The authors discussed intensity in Study 1 as if it were a manipulated variable rather than an observed variable. The use of observed emotion intensity is a substantive deviation from the Sheppes paradigm, limiting the extent to which results from the two can be compared.

**Our response**: If intensity were a product. We don’t observe a negative association. We observe a positive one but it does not pass standards of statistical significance. We aren’t attempting to explain study 1 with study 2. I state in the fucking article they are incomparable. Study 3 is the closest we come to a direct replication. Study 2 is not a conceptual replication.

1. Because this study does not attempt to manipulate emotion, we cannot know whether self-reported emotion intensity captured post-exposure is a product of or precursor to self-regulation.
2. A null association is inherently ambiguous. The results of the supplementary analyses contained within could be explained by emotion as either a product or precursor. Therefore, our conclusions may be overstated.

**Our response:** If Figure 4 represents emotion intensity as a product rather than a precursor, then

We do not overstate the null effect, but we accurately report our supplementary results. They are not meant to be taken as nullifying the original results. I think we are very clear that we are extending that paradigm to places it hasn’t been before. WE make no claims that the original hypotjesis wouldn’t be observed with trained subjects or less intense circumstances. Just that we observed this pattern in this context.

**Our response:** Why is it very likely? It’s perhaps equally as likely, but I have no reason to believe that it is more likely. There are two dissociable components to this critique. We agree that the analysis that we used did not manipulate and did not seek to manipulate emotions. Why would manipulation of an emotion differ from a non-manipulated one? They make more assumptions than we do, as they often use the standardized measure of emotion intensity rather than the experienced emotion. Though not perfect, some degree of affective manipulation did occur in the study. Subjects traversed the haunted house through 4 sections: two of which were deemed high intensity by the designers and two of which were deemed to be of lower intensity. When we analyze regulation behavior using these metrics, we still find the same regulatory behaviors.

Acknowledge that one means of accomplishing this could have been to have independent raters quantify the intensity of each event and to use those. This was difficult in our case given the time scale we had to work on.

Make the point that by taking an observational approach and capturing personal experience, we're forced to assume that the personal experience that they report is pre-regulation. However, if we were to take an IAPs approach, we'd also be forced to assume that the standardized value as we have represent personal experience.  
  
It is true that intensity could be either a product of or precursor to self-regulation in this study. While we argue that the language that we used (and "did you attempt to reduce or intensify this emotion?") suggests that we intended to target the emotional precursor (i.e., you may be able to prepare regulatory responses, but you cannot regulate something that hasn’t happened).   
  
. We [will change language in the manuscript to reflect this].

Specker et al., 2023 found that pre- and post-regulation emotion ratings were highly correlated (r = 0.81, p < 0.001). As best as I can tell, no one has tried to assess whether post-regulation intensity is a significant predictor of affective intensity. I could not find any articles which have explored whether there are still statistically significant differences in the post-regulation intensity of high and low intensity images after regulation. Shafir et al. 2016 did find a main effect of emotion category upon attenuation with a considerable effect size (ηp2 = 0.83), but this was calculated as the difference between the self-reported post-regulation rating and the standardized rating, so the baseline may be somewhat unreliable for this purpose.   
  
I tried searching for an emotion regulation dataset in which to test this theory, but none were publicly available.

**Our response:** It’s not intended to be read as the opposite, but it is intended to document the same phenomena. If the only want to comment upon Sheppes’s reappraisal and distraction is by using standardized stimuli in controlled settings with trained individuals whose experience we can dissect then no naturalistic study will ever comment upon them. We will be doomed to forever entertain a pristine, bottled version of emotion regulation that psychologists fawn over and the kind of self-regulation that everyone else actually has to use.

**Our response:**

**Our response:** I think it’s a large assumption to assume that this study design manipulates intensity. We’re just telling them intensity was. We can’t say Study 1 doesn’t match Sheppes but Study 2 does just because Study 2’s conclusions better fit Sheppes if we squint. There are many aspects of Study 2 that still differ from Sheppes, such that I would never call it a straight forward replication. It’s simply more similar in the specific dimension of controllability.

**Our response:** Make the point that studies 2 and 3 were not straight forward Sheppes replications. we get some patterns that are consistent with sheppes, specifically in participants that are either extremely psychologically distant from the experience (study 2) or making predictions about what should be normatively done for others (study 3). Emphasize the ways in which StuThe none option, measuring intensity before and after, using independently assessed intensity to predict self-regulationdy 3 deviated pretty substantially from Sheppes in both design and outcome.

Examine whether the other study that used individual intensity ratings also deviated from Sheppes [[SPECKER 2023]]

Claimed to be the first to measure pre- and post- regulation intensity in order to assess the importance of regulation flexibility (Specker 2023). Had Sheppes as a middle author. Their language is very confusing, but I think what happened is they aggregated trials and found, across the sample, an association between intensity and choice, but when they disaggregated, they found no association between intensity and choice. They kind of dismiss the latter finding as not having enough experimental control.

One possibility is that although individuals from clinical groups can use reappraisal successfully when cued, they fail to appropriately identify moments at which ER would be helpful in everyday life. Alternatively, members of clinical groups may in fact be able to identify moments at which ER would be helpful but for one reason or another choose not to use reappraisal very frequently in everyday life (as implied by Dryman & Heimberg, 2018). It is also possible that laboratory use measures capacity, which is an overestimate of actual success everyday life. Finally, it is possible that this disconnect between reappraisal use and success is an artifact of the way these constructs are measured (cumulative emotion ratings on a laboratory task vs. self-report responses). The source of the disconnect could have important implications for ER intervention science, which would respectively focus on using reminders and encouragement to select and initiate reappraisal in everyday life, improving conditions for implementing reappraisal in everyday life, or developing, refining, or combining measures of reappraisal that most closely correspond to documented emotional difficulties in clinical groups.

**Reviewer 4 Feedback:**

No comments