**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. 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 and attending to the effects of this difference is a central goal of this manuscript. 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 not identify any language that, to us, seemed to imply an inaccurate amount of experimental control was present, 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’. We have also added a statement acknowledging that having access to standardized ratings would have been an excellent supplementary opportunity and reasons why we weren’t able to achieve that on page 53, though, we do still contend that using standardized ratings alone would not be in line with the larger goals of this project. 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. 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 trying to capture 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 (See pg. 27 for details). 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 fundamentally change the structure of the relationship reported elsewhere in the literature between strategies and their success at different levels of pre-regulation intensity.

Distraction is often found to be more successful than reappraisal at attenuating emotions when the pre-regulation intensity is high. The opposite is true when pre-regulation intensity is low (Shafir et al., 2016; Sauer et al., 2016; Specker et al., 2024). High intensity emotions are less likely to be regulated successfully, regardless of strategy, (Sauer et al., 2016; Specker et al., 2024) so we should expect to see a negative pattern between success and pre-regulation intensity, regardless of strategy. It is also true that successfully regulated emotions attenuate (Shafir et al., 2016; Specker et al., 2024) so we should expect to see a negative pattern between success and intensity, regardless of whether it reflects pre- or post-regulation intensity. However, the attenuation has never been documented to be so strong that pre-regulation high-intensity emotions are regularly of a lower intensity than low-intensity scores post-regulation. The previously cited Specker paper and supporting literature within that manuscript have found pre- and post-regulation intensity reports are highly correlated (r = 0.809, p < 0.001). High intensity emotions do show greater reductions than lower intensity emotions (e.g., Shafir et al., 2015; Shafir et al., 2016; Szczygiel & Baryla, 2019), but, again, never so significant that they become less intense or indistinguishable from low pre-regulation intensity emotions. To the contrary, the difference in intensity of high and low intensity events seem to be significantly dissociable even when measured after regulation (from Specker et al.: t(127) = -25.9, p <0.001 ; this relationship was not assessed/reported by any other study as best as we are aware and these other studies almost always lacked pre-regulation self-reports of intensity to adequately contextualize such a finding).

All of this is to say, whether subjects reported pre- or post-regulation intensities, it would still be unusual in the context of the extant literature to observe reappraisal to be more successful than distraction at high intensities. In considering this critique, we did find some of our language regarding the interpretation on page 29 to be less careful than we intended and have amended it. We also added the results of an additional minor analysis (a respecified version of the simple slopes model with intensity as the moderator) to hopefully make our intended point more clear. We continue to emphasize that we have no evidence to suggest that this generalizes beyond the highly specific features of this context or without significant experimental manipulation and this evidence in itself does not conclusively explain what we observed within this context (pg. 30). As such, we have retained our initial interpretation of this result but have also added language to reinforce the reviewer’s concern that this method may be confused for an experimentally- controlled design.

*“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:** We agree that this was incorrectly stated. Distraction should not be interpreted to grow in efficacy as intensity increases (some events may be too intense to regulate at all), but it is relatively more efficacious than reappraisal as high intensities. As previously noted, we corrected this language, but also contextualized our finding by reminding the reader that there are methodological differences between our approach and the approaches that we are benchmarking ourselves against.

*“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:** Mixed interpretations of our question could potentially obfuscate a true effect, but such confusion would likely occur on the subject-level (i.e., different subjects may interpret the question differently, but the same subject would likely interpret the question the same every time they answer it). This sort of idiosyncratic effect is precisely what statistical approaches such as the hierarchical modeling we used are designed to adjust for. We agree with the reviewer that “… If emotional intensity drives use of distraction vs. reappraisal, we'd expect a positive association between intensity and distraction use” but feel that the logic behind “ … if [lower] emotional intensity is also the \*outcome\* of successful regulation … we'd expect a negative association between intensity and distraction use” is making the same mistake noted earlier: that distraction is so effective at attenuation that post-regulation intensity of a pre-regulation high intensity stimulus would look indistinguishable, if not lower, than pre-regulation low intensity stimuli. I believe that a more accurate expectation, based upon the cited literature, might be a positive association with a lower intercept or smaller logistic coefficient. We lack the ability to conclusively explain the null with this study design.

As we view it, the value of reporting this null is simply that when assess these variables in an ecologically valid way (i.e., minimizing manipulation) and in a context with features (i.e., highly stimulating, high intensity, complex) which mirror other circumstances in which self-regulation could be of vital importance, we do not find this relationship, despite an impressively substantial effect size and consistent replication in more controlled contexts. In considering this point, we made modifications to our abstract and significance statement to more accurately reflect that how we adjusted experimental control is likely an important component to what we observed across all studies; not just Study 1.

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

**Our response:** Study 2 was intended to more closely mirror aspects of the structure of a traditional Sheppes design and we expected to find the canonical relationship because of these additional experimental constraints. The intention was to demonstrate that conceptually similar but less-complex versions of the same stimuli could elicit the expected patterns, albeit much weaker than traditionally documented, if greater experimental control was added. We do not intend for Study 2 to explain the results of Study 1 (e.g., “*The different results observed in these studies are difficult to compare, though, as many features differ between the approaches…*”, pg. 38), but for them to be considered in parallel and as complementary to one another. The comparison is imperfect (hence the need for Study 3), but, we argue, still valuable as a demonstration that manipulation is, in part, essential to the existence of this effect, which conflicts with Study 1’s priority of external validity. This concern was noted by Reviewer 3 as well and our previous correspondence may be of relevance if Reviewer 1 is interested.

*“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). 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:** It was reasonable to replicate the Sheppes results for Study 3 given the level of experimental control. The intention was never to “disprove” the effect or claim it to be false, but to again illustrate the boundaries of observing the effect by altering aspects of experimental control (which is why we frame this as an “extension of previous work”). I think there might be a slight misunderstanding fueled by the mispredictions I made in the preregistration and how I discussed in on page 39 (which I have amended to hopefully make more clear).

Study 3 was motivated by Reviewer 3’s request to try to more directly compare Studies 1 and 2. In an ideal world, Study 3 would have been conducted in another quasi-naturalistic setting (ideally the same Study 1 setting) with some relatively more controlled design decisions (“*Because we could not incorporate an immersive experiential component such as in Study 1 in this experimental design*…” pg. 39), but timing and resources made this impossible. Our team decided to use video stimuli instead and to contrast forecasting and usage. The theory was that regulation decisions made within the idealized simulations people engage in when forecasting (which represents almost an additional level of control) should more closely resemble the canonical relationships documented elsewhere, while actual usage and the realities that complicate regulation would make the relationship between intensity and strategy ‘messier’, but I wasn’t sure whether it would be enough to dissipate the effect. I was confident that the videos would constitute a more complex set of stimuli than some comparable emotion regulation studies and that some design decisions (e.g., making the regulatory decisions non-binary) should reduce control and thus the effect, but I was not sure that it would be substantial enough to completely mitigate it. Subjects still received some training in identifying strategies, we still captured a relatively impoverished representation of the regulatory decision-making subjects may have engaged in, etc.

Ultimately, I predicted the statistical significance of the effect incorrectly, as noted in the pre-registration, but we still found notable differences in line with Studies 1 and 2 which contradicts expectations set by the existing literature. The categorical intensity of our stimuli were perfectly balanced between high and low intensity and yet we observed reappraisal to be used more often than forecasted rather than also perfectly balanced, as we saw among forecasters. We also found reappraisal to reduce emotion more than distraction reduces emotion, and that distraction was less reduced emotion less than subjects forecasted it would. We have made revisions to our discussion of these results to emphasize that Studies 2 and 3 contained greater experimental control than Study 1.

**Reviewer 4 Feedback:**

No comments