Reviewer 1 Feedback:

We appreciate the high-level feedback provided by Reviewer 1 and have provided responses to each critique. In some cases, we were not sure that we had interpreted the reviewer's request as intended. In these instances, we start our response with our interpretation of the critique. All following references to page numbers will refer to the clean version of the updated 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, raises a concern that we have incorrectly framed our predictor variable of interest as experimentally manipulated rather than observed. We thank the reviewer for bringing up this important point, and for helping us to clarify how our work contributes to existing emotion regulation research. In response, we have modified language throughout the manuscript to further delineate our work and distinguish it from Sheppes's work where appropriate. However, we have conducted additional post-hoc analyses using predictors from the study design that more closely mirror the original Sheppes work (this is discussed in more detail in the next response).
>
> We make comparisons to "Sheppes['s] foundational work" not because we believe that standardized ratings represent a direct comparison to idiographic self-report metrics, but because, conceptually, it is the preeminent standard to which almost all emotion intensity and regulation strategy research is currently being contrasted to. Few other studies in the extant literature, as recently noted by Specker, Sheppes, and Nickerson (2024), use subjects' self-reported emotional experiences as predictors of emotion regulation. Thus, the Sheppes paradigms remain the closest conceptual contrast to contextualize our results, even if we deviate from one another regarding experimental control of emotional intensity. 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). Our research primarily differs from prior work in the prioritization external and internal validity when modeling the target phenomena and attending to the effects of this difference. This was a central goal of this manuscript, and we therefore think that the comparison between this and prior work is valid and important.
>
> Confusion about experimental control was a concern cited by previous reviewers (See our exchanges with Reviewer 3). We had previously modified our manuscript to reflect that the relationship in Study 1 was not experimentally manipulated (e.g., "*It must be noted that we did not directly manipulate emotional intensity within this design*…", pg. 53) and to avoid language suggesting causality or experimental control ("*Study 1 tested whether the emotional intensity of negatively-valenced events was associated with the likelihood …*", pg. 10). In response to this critique, we have again carefully reviewed the language we used to describe Study 1. To ensure that we are not giving the impression that our predictor in Study 1 was experimentally manipulated, we have added additional statements on pages 5 to 10, 20 to 21, 26 to 28, 30 to 32, 39, and 53. We have also added an explicit statement acknowledging that having access to standardized ratings would have substantially strengthened the contribution of the present research, and note reasons why we weren't able to get these ratings in our present design on page 53. There may still be some examples of inaccurate language, which 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 also an important point that was not previously discussed. 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 intended to target emotion as a precursor and chose language trying to capture it as such, affective or regulatory assessments may have been biased or

misreported. If subjects did interpret our post-exposure questions (i.e., how intense was this experience) in different ways or if self-report values were in some way a representation of both pre- and post-regulation intensity, it may have had an effect upon the relationship between strategy usage and emotional intensity. There are two relevant pieces of information which may mitigate this concern: Specker et al., 2024 – one of the only studies to assess emotion intensity before and after regulation using distraction or reappraisal - found that pre- and post-regulation intensity scores were highly correlated ($r = 0.809$, $p < 0.001$). Additionally, we used hierarchical modeling to account for idiosyncratic subject-level effects and we believe that confusion or bias between pre- and post-regulation intensity 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). Regardless, we have updated our text on pages 20 and 53 to reflect this limitation and addressing it has improved the manuscript. 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, we have conducted two additional analyses. The first examines 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 confirmed the intended design structure (See pgs. 15, 26 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. The second analysis computes an aggregate rating of intensity for each unique observation using a leave-one-out approach. This attempts to predict regulation behavior using relatively more normalized representations of the emotion intensity of each event. With this technique, the predictor becomes the average intensity reported by all ***other*** subjects (i.e., excluding the 'experiencer's' self-reported intensity) experiencing that same event in each observation. This is a higher resolution approach than using section to predict regulation, but, again, no observed association between intensity and usage was observed. These analyses are of course imperfect, post-hoc solutions to this criticism, but we believe they at least add credence to the notion that our null results in Study 1 are not simply a product of lack of measurement control. These changes are covered on pages 20 and 26 to 28.

*"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: We believe that the previously noted post-hoc analysis does address this precursor/product concern to some degree, but we would also like to emphasize that the result of interest in Figure 4 is not the association between intensity and success, but that at most intensity values reappraisal was reported as more successful than distraction. Finding reappraisal to be a more successful strategy at higher intensities is atypical in the context of the extant literature (Shafir et al., 2016; Sauer et al., 2016; Specker et al., 2024) whether subjects reported pre- or post-regulation intensities. We included the results of an additional analysis - a respecified version of the simple slopes model with intensity as the moderator - to hopefully make our intended point clearer. We do deviate from the previously cited work in that this success metric is subjectively assessed; we do not have pre- and post-regulation intensity to compare, which would be a relatively more objective means of measuring this. We have updated pg. 30 to reflect that. In considering this critique, we reworded our interpretation of this result. Page 30 now further emphasizes that: 1) we have no evidence to suggest that this generalizes beyond the highly specific features of this context and 2) this evidence in itself does not conclusively explain what we observed within this context.

"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 with the reviewer that this was incorrectly stated and thank them for bringing this to our attention. 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 at high intensities. As previously noted, we corrected this language, but also contextualized our finding by reminding the reader that there are important 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:** We interpret the reviewer's response to suggest that we should expect an interaction between differences in pre- or post-regulation intensity by the strategy used. In other words, if the same high intensity event were to be regulated by two identical people, one using distraction and one using reappraisal, we should expect distraction to produce a lower post-regulation intensity than reappraisal. If responding to an identical low intensity stimulus, we might expect the opposite: a lower post-regulation intensity for reappraisal relative to distraction. Because our design cannot separate pre- and post-regulation intensity, instances in which subjects report pre-regulation intensity might offset instances in which subjects report post-regulation intensity, resulting in a null effect. If we have misunderstood, please let us know.
>
> If interpreted correctly, this seems to be describing the relationship between strategy *effectiveness* at different emotion intensities. This is undoubtedly a critically important question to answer in naturalistic settings, but one that we think is outside the scope of the present work. The present work focuses on strategy usage *likelihood* at different emotion intensities. Fully addressing this point would likely require running a new set of studies wherein this question is the primary focus, more standardized stimuli that could be repeatedly used over time, the controlled and equal use of multiple strategies across participants, and potentially a more extensive longitudinal design.
>
> We do lack the ability to conclusively explain the null with this study design and acknowledge that. As we view it, the value of reporting this null is simply that when assessing these variables in an ecologically valid way (i.e., minimizing manipulation) and in a context with features which mirror other circumstances in which self-regulation could be of vital importance (i.e., highly stimulating, high intensity, complex), we do not find this relationship. This is despite an impressively substantial effect size and consistent replication in more controlled contexts. As such, we believe that the present research helps to identify boundary conditions which should be more iteratively tested. Studies 2 and 3 contribute somewhat to this goal, but are far from exhaustive. In considering this point, we have 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. We expected to find the canonical relationship because of these additional experimental constraints. We have made this expectation clearer in the present manuscript (pgs 31 to 32) to reduce confusion. We intended to demonstrate that conceptually similar but less-complex versions of the same stimuli could elicit the expected patterns. 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, as complementary to one another, and to situate the present work in the context of the extant work on regulatory strategy choice/usage. 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.

"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: The goal of Study 3, as well as Study 2, was to conceptually replicate Sheppes's original work rather than to "disprove" the effect. We intended to illustrate the boundaries of observing the effect by altering aspects of experimental control. We frame this as an "extension of previous work" and have changed how we discuss the study goals to further reiterate this point on page 39.

Notably, the design of 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 (the haunted house only runs one month of the year) and resources made this impossible. Our team decided to use video stimuli instead and to contrast forecasting and usage. The hypothesis 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', perhaps not enough to completely dissipate the effect between intensity and usage but to mitigate it.

Ultimately, our predictions were not supported by the results of Study 3, as noted on page 39. However, we still found notable differences in line with Studies 1 and 2 which contradict expectations set by the existing literature. The categorical intensity of our stimuli was balanced between high and low intensity and yet we observed reappraisal to be used more often than was forecasted. We also found reappraisal to reduce negative emotion intensity more than distraction reduces negative emotion intensity, and that distraction reduced negative emotion less than subjects forecasted it would. We have revised our discussion of these results to emphasize that Studies 2 and 3 contained greater experimental control than Study 1.