

Emotion Regulation Capacity Lacks a Shared Neural Signature

Corresponding Author: Dr Maurizio Sicorello

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

Sicorello and colleagues applied a rigorous and preregistered meta-analytic approach of 40 fMRI datasets to elucidate the association of three measures of the emotion regulation strategy „reappraisal“: task-based affective self-report ratings, amygdala downregulation, habitual/trait self-report questionnaires.

The article is well written and the methods are sound. Given the importance of the reappraisal construct in psychology and mental health research, the critical question of how self-report and brain activity are associated that reaches beyond emotion regulation research and the expandable, innovative methodological approach taken by the authors, this manuscript does have the potential to impact the field. This speaks in favor of a publication in an outlet like Nature.

We do have several comments that can principally be addressed regarding the theoretical rational of the study and the conclusions drawn from the data. Some minor points are listed at the end of this review.

Our main question concerns how the overarching research objective is framed. The authors aim to find a common core on which different measures of reappraisal converge and interpret the lack of such convergence such that brain activity in reappraisal tasks should not be interpreted as indicating individual reappraisal capabilities. While the authors elaborate on different perspectives, this is, in a nutshell the main message. In our view, this gives trait self-reports an unjustified weight as „ground truth“ against which other measures must prove themselves. Rather, conceptually, each measure assessed reappraisal but targets specific aspects of the reappraisal construct: (1) Trait questionnaires mainly assess the beliefs that people have about themselves regarding reappraisal. (2) Task-based ratings are momentary snapshots and, thus, reflect the effect of reappraisal on current affect more directly, but are influenced by the ability of participants to access their current affective state and report on it. (3) Neural responses are not influenced by participants' beliefs about themselves or the subjective accessibility of affective states. In that sense, they could be seen as closest an organism's reappraisal capabilities. However, the summed BOLD activity assessed with fMRI is, of course, only a fragment of the brain's actual functionality. An additional dimension that the three measures may vary on is whether they assess the propensity to engage reappraisal (rather strong in trait self-reports) vs. the capacity to which this can be done (rather strong in tasks instructing participants to engage reappraisal). We think that the manuscript could be stronger and (even) more interesting to the field if it aimed to make use of the massive amount of data they accumulated to bring out, explore and explain these differences. Future studies could then evaluate what measures to include to assess, for instance, deficits in patient groups.

Relatedly, the authors argue that the lack of an association of the three measures is due to them being interindividual associations and the methods not possessing enough interindividual variability partly because they were not constructed for assessing that variability. The fantastic chance that the authors have is to actually test that interpretation for two of the measures, brain activity and task-related affect ratings, as these were acquired simultaneously across the duration of the experiment. So intra- and interindividual variation is available and could be compared. We understand that this would require reanalysis of the first level fMRI data together with the ratings but believe it would be an important step.

The authors discuss in detail the possibility of overall interindividual differences in brain activity „overshadowing“ any

reappraisal related differences and thereby preventing significant brain-behavioral correlations. What did not become clear to us is intraindividual contrast images (e.g. reappraisal vs. view) do not control for this? Does the data indicate an interaction of overall activation differences with experimental condition?

Minor:

L. 128, first sentence: error in Referencing

L. 138f. Reappraise vs. permit emotion – is “permit emotion” the “watch” condition?

L. 238 Jackknife procedure: Why do both studies have a large impact? Is it because of their sample size? Maybe turn the paragraph around and say it in the beginning. This way the reader knows in the beginning why they are influential

L. 246f. Is the neurosynth analysis presented anywhere in the paper?

L. 281 reappraise vs. view – maybe stick with this term instead of “permit emotion” which sounds unfamiliar to me and inhibits reading flow

L. 281f. “only negative signs” – maybe one sentence about the meaning here?

L. 289 bracket missing

Correlations have been indicated with two numbers after the decimal point – maybe this is due to the journal's statistic reporting guide?

(Remarks on code availability)

Referee #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Referee #3

(Remarks to the Author)

This study addresses the construct validity of “reappraisal capability” by examining the interrelationships between three hypothesized ways of measuring it, including trait questionnaires of habitual reappraisal, reduced self-reported negative affect when presented with negative stimuli in tasks, and reduced amygdala activity when presented with negative stimuli in tasks.

The immense amount of collaboration required to do this project is laudable. All teams used a uniform set of procedures to calculate intercorrelations between trait, self-report, and neural measures. For the analytical approach, the authors calculated correlations between measures across individuals for each study separately. They then used these statistics in subsequent analyses. This approach is distinct from more recent “mega-analysis” approaches where a single model is used to combine data across sites and studies at the subject level, but it is reasonable given the heterogeneity of the datasets and measures. Regardless, the volume of data and the approach is certainly sufficient to test with confidence whether there is support for even a small correlation at the population level (see point 7 below for additional support). Additional analyses examined other brain regions and also correlations between brain regions in this data and also in another multi-task dataset, that more generally challenge the assumptions of the conventional “up/down” regulation model that underlies the neuroscience of emotion regulation.

This is an exceptional paper that will undoubtedly generate a great deal of discussion, debate, and impact in the field. It is well-executed and contributes field-challenging conclusions that researchers will have to reckon with. The manuscript is clearly and concisely written. I only have minor comments for the authors below.

Minor:

1 Given the lack of convergence between measures, the authors' main suggestion for future work is that the field needs better alignment between neural measures and constructs. As they note in the discussion section: “The weak convergence between common outcome measures and their inconsistent associations with canonical reappraisal networks underscore the need for better-aligned behavioral, trait-based, and neural constructs for them to capture a shared core of cognitive reappraisal.” Here, the authors could also consider that the theories and statistical approaches that underlie this work assume relatively simple mappings: e.g., dominantly linear or at least monotonic functions, and one-to-one mappings between brain regions/networks and psychological phenomena. Their methodological approach makes these same assumptions, even if it is to show that this approach ultimately leads to challenges in construct validity. But there is also an alternative future direction. Increasing work in affective neuroscience suggests that the mind–brain mapping might be more complex (i.e., may involve degeneracy or multiple solutions). By this account, refinement of measures may still not lead to clear relationships. Rather, alternative theoretical and analytical approaches are required that account for complex mappings. Note, this is not a criticism of the authors' work—but rather a comment that could be discussed in the review process only or could be considered by the authors in a revision to stated future directions.

2 I think the results section could use a bit more clarification on the measures they are using for statistical analysis. Information from this sentence from the methods section would be helpful to reiterate or paraphrase in the first subsection of

the results section: “Research groups prepared up to three unthresholded group-level statistical images, containing voxel-wise t-statistics for the between-person correlation of the whole-brain neural response in the [reappraise - view] contrast and the three outcomes of interest.”

3 “A person’s ability to either down-regulate their amygdala or up-regulate regulatory regions is often interpreted as an indicator of interindividual reappraisal capability” — consider rephrasing slightly since the current phrasing can be read in two different ways, one of which implies dualism.

4 This part of the results section could use a bit more clarification: “...people differ in their global brain responses, likely due to biological confounders, which might dominate any specific regional effects and dilute their usefulness as a marker of an individual’s emotion regulation capability.” In general, the “Amygdala down-regulation” section could be developed further to clarify the argument and the counterargument.

5 Is the positive correlation true across tasks for default mode and frontoparietal networks, too?

6 The “statistical power” section seems a bit detailed for the main text and could be moved to supplementary, and the main text revised to simply state that numerous approaches were performed to validate statistical power.

7 To further convince skeptical specialist readers that they have enough data, one additional and relatively simple analysis they could include is to generate data simulations assuming the 40 studies with their specific sample sizes and a true population correlation of .1 or so, and see if their simulated data rejects a null hypothesis. While I was convinced by the authors’ thoroughness already, I did this in addition and found it useful.

(Remarks on code availability)

Referee #4

(Remarks to the Author)

Overall comments

This paper combines data from over 40 different fMRI Studies of reappraisal to address questions about the extent to which behavioral and brain measure of reappraisal correlate with one another. Given that emotion regulation is a core ability central to healthy adaptive functioning, and that problems with regulation characterize nearly every common clinical disorder, this paper addresses an important topic.

That said, I do not believe this paper in its current form is ‘complete’, methodologically and conceptually thorough and suitable for publication. The questions the paper seeks to address - about the nature and correspondence of behavioral and neural measures of reappraisal as a paradigmatic emotion regulation strategy - are important *and* very complex. As such, addressing them appropriately demands a great deal of context and nuance - both analytically and - perhaps as critically - conceptually, when discussing the core psychological constructs and neural systems of interest.

The lack of nuance/consideration of complex analytic and conceptual issues muddles the paper’s message and potential impact. The paper reports largely null results, mentions potential reasons for this such as commonly used reappraisal tasks not being ecologically valid, and notes that there may also be widespread issues in fMRI research like multiple regions showing high degrees of positive correlation (i.e. inter-regional comparisons are confounded by so-called global signal issues). The paper does not, however, unpack what any of this means. This, I believe, is absolutely crucial. A paper of this kind must, I believe, consider the ways in which:

(Point a) the ways in which the current analytic approach - which relies entirely on study-level correlation maps - might be limited in terms of addressing the core issues.

(Point b) the conceptual complexity of the core issues the paper purports to tackle - ranting from a theoretical analysis of why one would or would not expect different behavioral and neural measures of reappraisal to correlate to what it means to talk about emotion regulation capability.

(Point c) the design / approach of current reappraisal studies may or may not contribute to the apparent null findings (e.g. is this an issue of approach / method - could future studies be different, e.g. more ‘ecologically’ valid?). As opposed, that is., to the point b), where the issue is that different behavioral and neural ways of assessing regulation really would not be expected to relate to one another, based on a first principles analyses of what different measures reflect?

Detailed comments

(Point a)

With respect to core questions about the current analytic approach:

— A key question is whether the current analytic approach is the appropriate one for addressing the core questions at hand about brain correlates or reappraisal corresponding to behavioral measures of reappraisal. The present approach states that

is takes account of an N of over 2000 going into the main analyses, which involve the aggregation of study-level correlation maps for 41 fMRI Studies of reappraisal. Given that study level maps - not individual subject data - are entered into analyses - it does not seem appropriate to state that these analysis were based on an end of over 2000 participants. The real N is, in essence, the number of studies included in the paper, which is 41.

That said, the core conundrum is whether results would be different if the authors did their analyses differently.

First and foremost is the question of what the results might be if the authors focused not on study-level maps, but had access to the actual subject-level imaging data. This is important for many reasons. Perhaps chief among them is the possibility that truly meaningful individual differences in underlying neural dynamics might only be observable in subject-level data - in the ways the each individual relies on different combinations of key underlying neural systems when reappraising. By relying on study level maps, the current analysis is insensitive to such within individual differences. As such, they key sources of variability in the neural systems underlying reappraisal may be obscured by the averaging that goes into study-level maps.

Second, there is a question of why the authors focus only on changes in amygdala response rather than multivariate patterns reflecting changes in emotional experience. As described in other comments below, the amygdala is seen by a few effective neuroscientist today to be a strong direct contributor to emotional experience. Add Andersen and Elizabeth Phelps a very nice data from over 20 years ago showing that daily emotions reported by patients who lack the amygdala are pretty much the same as those found in healthy individual individuals. More recently, multi various neural signatures like the PINES signature from Chang et al, have been validated to correspond to variations and self-reported negative affective experience elicited by the kinds of stimuli used in reappraisal studies. You can take the amygdala out of those neural signature maps and it doesn't change their ability to predict experience. Did the authors consider doing an analysis with the PINES or some other neural signature?

Third, the authors failed to see their purported indices of reappraisal capability corresponding with activity in any of the prefrontal systems associated with reappraisal identified in a meta-analysis. There have been recent papers publishing neural signatures associated with reappraisal. Have the authors considered, asking whether expression of these patterns might relate to changes in either self reported affect during reappraisal or in changes in the expression of neural signatures for effective responding of the sort noted in the prior paragraph?

(Point b)

From a conceptual perspective, the paper does not present clear hypotheses or a logic as to why one would - on the basis of theory or prior empirical results - one should or should not expect to see correspondence between the measures reappraisal included in the present paper.

— The paper could take greater care in its use of terms. For example, in the introduction it is stated that, “Compounding this issue, regions activated during reappraisal also subserve other functions, making it unclear whether they reflect regulation capabilities per se...”. This comment seems to confound a description of the function of a region with the behaviors/tasks during which it is activated. A region could have a singular ‘function’ - defined in terms of the specific computational processes it implements - and may participate in multiple different behaviors. Do the authors mean regions engaged during reappraisal are also engaged during other tasks, such as those involving working memory, response selection, etc? If so, it is possible that, for given region, one or more common underlying functional processes plays a key role in all of these behaviors.

— The paper talks about the fact that, “...task designs may not mirror naturalistic real-life regulation, raising questions about their validity and practical relevance for important individual differences...”. This is an idea shared by many in the field, and this may be an absolutely critical issue for understanding what the present results mean, but the present paper doesn't do much to explain why this might be the case. In what way are the tasks used in the scanner not picking up on factors that may determine emotional health and well-being in everyday life? The answer to this question - and hypotheses about it - may be essential for understanding what the present results mean.

— “fronto-limbic” is antiquated/outdated terminology. The concept of the limbic system is outdated. See numerous papers from Joe LeDoux (e.g. Neuron paper from ~2014) detailing why the idea of a limbic system as a functional ‘system’ is simply incorrect.

— At the outset, the paper states that, “We examined three widely used indices of reappraisal capability 25: (1) self-reported habitual reappraisal (trait questionnaires), (2) task-based reductions in negative affect (task-based affective ratings), and (3) task-based reductions in amygdala activation (amygdala down-regulation). This enabled us to ask whether these indices reflect a shared underlying core construct—and whether individual differences in reappraisal capability, as indexed by each measure, are systematically associated with neural activity in a priori defined regulatory neural networks.” Unpacking these statements is important for understanding some of the shortcomings of the present paper's construction/presentation.

These comments seem to miss a concept that is central to research on emotion, but is often unappreciated by those in affective neuroscience - the idea of coherence between different channels of emotional response. As described in reviews by Cacioppo and colleagues, and more recently by Mauss and Gross (and as unpacked in some nice empirical papers from Mauss's lab), emotional coherence refers to the correlation experiential, behavioral and physiological channels of emotional responding. It turns out that ~50% of the time, these channels don't cohere. That is, experiential self-reports, physiological

and/ behavioral measures (e.g. facial expressions) only correlate with one another about 50% of the time. There seem to be numerous reasons why this might be the case (e.g. different time scales for different measurements, differing processing systems underlying each measure, different covariates having differential influence over each measure etc.).

Against this backdrop, it could be very interesting to frame the present paper as introducing a different version of the coherence question - do so-called trait measures of reappraisal, self-report-based indices to task-based success and measures of activity in brain systems intercorrelate? Do they cohere? To properly unpack this question, the present paper would need to present a theoretical analysis of why we would or would not expect these measures to correlate/cohere - akin to what has been done in the broader coherence literature. What does measure reflect? To what extent are they thought to reflect the operation of similar vs. different underlying processes/mechanism? Etc. The present paper does not do this.

— Wrapped up in this issue seems to be an assumption that reappraisal 'capability' is a clearly defined psychological concept/construct. At different points the paper treats scores on self-report questionnaires of reappraisal tendencies as well as reappraisal-related changes in task-based self-reports of emotion and brain activity as indicators of 'capability.' For example, in the discussion the authors write, "These findings suggest that the three common indices of reappraisal capability capture qualitatively different facets of reappraisal, despite often being interpreted interchangeably. The absence of a common core among these indicators challenges the assumption that they reflect a shared underlying capacity."

Here it would be essential for the authors to define what 'capability' means. The ERQ is typically seen as measuring beliefs about 'tendencies' to engage in reappraisal. Changes in self-reports of affect might measure capability - in a lab setting, when the experimenter tells people when and how to regulate. And changes in amygdala response - which once might have been taken to reflect changes in 'emotional response', broadly construed, are now not seen by many to be that. Instead, the amygdala is now seen more as a neuromodulatory region that influences the focus of attention, speed and fidelity of perception and consolidation of information into memory. Few affective neuroscientists today think its activity 'should' correlate with self-reports of emotion. Instead, current approaches look to multivariate approaches to define neural signatures that have been validated to correspond to variations in self-reported emotional states (e.g. the PINES signature from Chang et al).

With this in mind, finding such as, "...participants with stronger down-regulation of the amygdala via cognitive reappraisal also showed a better down-regulation of task-based affective ratings. Yet, the effect size was comparatively small...." don't really land with this reader, because they seem to depend on a kind of 'straw man' set of naive assumptions about what these different measures reflect.

— Along these lines, the authors might want to think about whether there is something new and interesting their data can say about so-called trait tendencies vs. task-based measures. For example, the authors report that, "There was no significant association between trait questionnaires and either outcome, i.e., amygdala down-regulation or task-based affective ratings." And that, "There was no statistically significant association between trait questionnaires and neural activity during reappraisal."

On one hand, it would be great to have these findings published in the literature. They address a file-drawer problem for work on emotion regulation where many who have done reappraisal studies - and have included the ERQ - have known for years that ERQ scores do not predict either changes in task-based changes in self-reports of affect or changes in amygdala or other brain correlates of reappraisal. To date, however, no one has published these failures to see such correlations.

On the other hand, once noted, the key question is WHY 'trait' measures, like the ERQ, predict good outcomes in everyday life but do NOT correlate with task based measures of reappraisal-related changes in emotion. The present analyses and discussion do not provide an answer. One way to look for answer could involve drawing on variants of 'extended process models' of emotion regulation (see, e.g. examples from Gross, Sheppes, Ochsner and colleagues) positing that the implementation of a strategy like reappraisal comes only after a series of other processing stages have occurred. These stages include, critically, an initial decision about whether reappraisal (or some other strategy) is called for at the present moment, and then if regulation is to happen, a decision about what strategy to use. It is very likely that 'trait' measures like the ERQ pick up individual variability in these decision/selection processes as much, if not more, than they pick up on variability in the fidelity with which individuals can implement specific reappraisals when told when and how to do so. The latter stage - implementation - all that current fMRI studies tap into.

— How do the authors make sense of their failure to see any PFC regions associated with individual differences in light of evidence collected in fTMS and neurofeedback studies that individual differences in PFC engagement during reappraisal are important? For example, recent studies of reappraisal in clinical groups seems to show that ventrolateral PFC activity during reappraisal is weaker (De la Pena-Arteaga et al., 2021) and may inversely correlate with symptom severity in MDD and BPD (Keller et al., 2022). Notably, however, in patients with MDD, neurofeedback focused on ventrolateral PFC can increase the effective daily use of reappraisal (Keller et al., 2021). See also recent paper by He et al in J. Neuroscience (2023).

(Point C).

The paper mentions that currently used scanner-based appraisal tasks may not map onto what people do in everyday life when regulating emotion. That is, current tasks may not be ecologically valid. This is a very interesting idea. A likely very important one. Can the authors say more about what they mean and how this might bear on the findings they observed?

— The above discussion about extended process models might provide some insights on this issue. For example, if current fMRI tasks only measure the implementation of a strategy in a highly constrained context where individuals are told when and how to re-appraise, and are given practice in doing so, then such procedures may get rid of many sources of naturally occurring individual variability that are very meaningful to the full range of individual differences in emotional well-being the authors are interested in. These include factors like those mentioned earlier, including the factors that determine whether an individual chooses to regulate, and if so, how.

— Another consideration here - and one that bears on the question of coherence across trait versus task-based measures of reappraisal, comes from recent work by Michael Inzlicht and colleagues. Inzlicht has noted that “trait” measures of self-control - that is, self-report questionnaires measuring some aspect of self-control - and “state” measures of self-control - that is, task-based measures of the ability to exert a specific kind of self-control - are often related to different kinds of outcomes. Notably, he argues that the trait-based questionnaires are what predict most of the good things we typically think of when we think about self-control, whereas the task-based measures rarely predict anything terribly meaningful. See his recent review/opinion papers on this. Could this illuminate a discussion of the short-comings of current task-based approaches for studying reappraisal?

— is there any potential issue associated with the fact that nearly of the 41 reappraisal studies included in the current analysis used photographs as their stimuli? Assuming most of these come from the international affective picture system, or some similar set, then these images were created to as to elicit affective responses that are normatively the same across individuals. Could the use of such stimuli actually work against finding meaningful individual differences?

(Remarks on code availability)

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)

The authors have done a great job in addressing the previous concerns. I have no further comments and believe the manuscript to be fit for publication.

(Remarks on code availability)

Referee #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Referee #3

(Remarks to the Author)

The authors thoroughly addressed my comments. I believe that this will be an impactful article in the field that will generate a great deal of discussion.

(Remarks on code availability)

The main concern I had was whether one of the analytical approach they used was adequate. They took lots of small N studies, calculated correlation maps in each of them, and then pooled across those correlation maps. I was concerned whether this approach had enough sensitivity to detect effects since it is unconventional, and so I ran my own simulation to check and found that the approach was sound. The authors now report including their own simulation - which clearly shows that they have enough power.

Referee #4

(Remarks to the Author)

I have to say I wasn't sure the authors could address the analytic and conceptual issues I saw in the paper the first go round..... but I'm happy to say they did. For this reader, the initial question wasn't whether this kind of paper would be valuable - but whether the way the paper was framed and the analyses presented in the initial submission did justice to the questions and issues at hand.

So. This revision is among the most comprehensive and in depth I have seen. I'm impressed with its thoroughness and thoughtfulness. The new analyses suggested by all the reviewers should significantly strengthen the paper and the new conceptual framing should help the work have a broader impact.

That said - there are two somewhat minor remaining conceptual niggles.

The first concerns the characterization of the ERQ. I know this measure well. It does not measure 'capability'. It measures self-reported tendencies to engage in specific patterns of thinking that characterize reappraisal. I do not believe it is appropriate to characterize it as a capability measure. The authors have revised the paper to provide a definition of capability that is broad enough to nod in this direction. I suppose I wish the authors did a bit more to explicitly acknowledge that many in the field see this as a tendency measure.

The second concerns how we might interpret some of what authors report - like ERQ scores positively predicting negative affect-related DVs. We tend to think that when we try regulating that all DVs reflecting affective responding should go down. This paper suggests that isn't always the case and that different DVs don't really track with each other. One wonders is two things could be doing on here. One is that reappraisal could lead people not to simply diminish emotion but to change it. To the extent that any DV reflects arousal/intensity of emotion rather than the quality/valence of emotion (e.g. the amygdala!) then we might not always expect its activity to go *down* when we reappraise. A second question is whether some people tend to try reappraisal, in everyday life or the lab, because they tend to have strong (negative) emotions. Their emotions might be so strong that - despite their efforts - they try and fail. The could be because the emotions were too strong. It could also be because reappraisal might not have been the best strategy to use then (See papers on reappraisal choice, e.g., or EMA Studies, suggesting we tend not to choose reappraisal when emotions are strong), or that they aren't very good at it (for whatever reason). There are examples in the literature (e.g. papers from Mauss's lab about reappraisal not always being the best thing to do), including clinical examples (e.g. recent paper from Herzog et al with depressed individuals) consistent with this idea.

Offering more food for thought about how to interpret some of what was found, including some of the results from the new analysis....

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response Letter

REVIEWER 1&2

Sicorello and colleagues applied a rigorous and preregistered meta-analytic approach of 40 fMRI datasets to elucidate the association of three measures of the emotion regulation strategy „reappraisal“: task-based affective self-report ratings, amygdala downregulation, habitual/trait self-report questionnaires.

The article is well written and the methods are sound. Given the importance of the reappraisal construct in psychology and mental health research, the critical question of how self-report and brain activity are associated that reaches beyond emotion regulation research and the expandable, innovative methodological approach taken by the authors, this manuscript does have the potential to impact the field. This speaks in favor of a publication in an outlet like Nature.

We do have several comments that can principally be addressed regarding the theoretical rationale of the study and the conclusions drawn from the data. Some minor points are listed at the end of this review.

R1-C1:

Our main question concerns how the overarching research objective is framed. The authors aim to find a common core on which different measures of reappraisal converge and interpret the lack of such convergence such that brain activity in reappraisal tasks should not be interpreted as indicating individual reappraisal capabilities. While the authors elaborate on different perspectives, this is, in a nutshell the main message. In our view, this gives trait self-reports an unjustified weight as „ground truth“ against which other measures must prove themselves. Rather, conceptually, each measure assessed reappraisal but targets specific aspects of the reappraisal construct: (1) Trait questionnaires mainly assess the beliefs that people have about themselves regarding reappraisal. (2) Task-based ratings are momentary snapshots and, thus, reflect the effect of reappraisal on current affect more directly, but are influenced by the ability of participants to access their current affective state and report on it. (3) Neural responses are not influenced by participants' beliefs about themselves or the subjective accessibility of affective states. In that sense, they could be seen as closest to an organism's reappraisal capabilities. However, the summed BOLD activity assessed with fMRI is, of course, only a fragment of the brain's actual functionality. An additional dimension that the three measures may vary on is whether they assess the propensity to engage reappraisal (rather strong in trait self-reports) vs. the capacity to which this can be done (rather strong in tasks instructing participants to engage reappraisal). We think that the manuscript could be stronger and (even) more interesting to

the field if it aimed to make use of the massive amount of data they accumulated to bring out, explore and explain these differences. Future studies could then evaluate what measures to include to assess, for instance, deficits in patient groups.

Response:

We thank the reviewers for this conceptually important reframing. We agree that each measure may capture a distinct facet of the reappraisal construct, and we wish to clarify that this was precisely the question our study was designed and powered to test. Our intention was not to position trait self-reports as privileged or as ground truth, which we now clarify more explicitly throughout the manuscript.

We had two related preregistered goals, whose equal importance we now emphasize more clearly in the Introduction: (1) to test the extent to which the three most widely used operationalizations of reappraisal capability — trait questionnaires, task-based affective change, and amygdala down-regulation — are associated with one another and converge on a shared underlying construct, and (2) to test whether each of these operationalizations is systematically associated with neural activity in canonical regulatory networks (e.g., prefrontal cortex), originally expecting a pattern of partially overlapping subregions.

The reviewers propose an insightful conceptual framework: trait questionnaires may primarily index beliefs and habitual tendencies, affective ratings may index momentary regulatory effects, and neural responses may index something closer to biological regulatory implementation, independent of self-report. We find this framework compelling and have incorporated it into the revised manuscript, with especially extensive changes to the discussion. Notably, three large studies with a total sample size of $N = 444$ were able to provide data for additional questionnaires, precisely, ERQ versions assessing beliefs about implementation ability rather than habits. Beliefs about and implementation ability of emotion regulation were very highly correlated ($r = .67$, 95% CI = [0.62, .72]). The analysis of association of brain activity from the emotion regulation tasks with beliefs about implementation ability also yielded non-significant results in this subsample. We believe this is evidence that the misalignment among the three analytic levels (trait questionnaires, task-based affective ratings, neural responses) is not only due to the habit-aspect of questionnaires like the ERQ, but rather reflects a broader dissociation of brain activity and self-reports.

The reviewers' suggestion that future work should characterize what each measure uniquely contributes is one we strongly endorse. We believe this will require external validation criteria that do not trivially overlap in modality with any of the three measures (e.g., another questionnaire), with clinical outcomes in diagnostically characterized transdiagnostic samples representing one especially promising direction. At the same time, we believe our second goal — relating these three indices to canonical regulation networks — is also informative in this context, as none of the three indices showed the hypothesized associations that are often assumed in the literature. As the reviewers suggested, we believe a natural next candidate is clinical outcomes in transdiagnostic samples, which represents the next long-term phase of the NIDER consortium.

We have revised the Introduction and Discussion to frame the three measures more explicitly as distinct but theoretically related operationalizations of a multicomponent construct, and to present the absence of convergence as an empirical finding that calls for the more differentiated interpretation the reviewers recommend.

In the revised Introduction, we tried to make brief clarifications, including core examples from the emotion regulation literature, suggesting a relation between these three levels:

“We focused on three widely used indices often assumed to reflect related facets of reappraisal capability: trait questionnaires, task-based affective ratings following reappraisal, and amygdala down-regulation during reappraisal. Prior work has linked these three indices with each other, and with prefrontal responses during reappraisal^{12,15–18}, suggesting a common neural substrate.”

The following statement of the goal aimed to introduce the different analytic levels with equal weight:

“We asked whether the three indices converge on a shared core, and whether any of them relate to activity in canonical emotion regulation circuits during common reappraisal tasks².”

The Discussion now opens with the coherence framing explicitly:

“The lack of convergence challenges a central assumption in affective neuroscience and clinical neuroimaging: that questionnaires, task-based affective change and neural responses provide complementary windows onto a shared underlying regulatory capability, albeit with unique aspects. This matters because the extent to which common reappraisal task findings inform claims about emotion, psychopathology and intervention targets may be overstated if individual indices capture only distinct aspects of regulatory functioning. Indeed, research on emotional coherence shows that experiential, behavioral and physiological channels often diverge because they operate on different time scales and are shaped by different causal influences and biases²⁷. Trait questionnaires reflect recollected beliefs about reappraisal habits and ability across diverse real-life contexts; task-based affective ratings capture momentary regulatory effects under constrained experimental conditions; and neural responses may additionally reflect processes outside conscious awareness²⁸, while being shaped by neurovascular and scanner-related factors unrelated to regulation. The present findings therefore do not merely show weak measurement convergence; they suggest that common indices of reappraisal capability, in fact, may carve the construct at different joints.”

Later in the discussion, we explored the likely reasons for this disconnect more extensively based on our present and independent previous findings:

“The absence of associations between trait questionnaires and task-based or neural indices points to an explanatory gap that can be understood through process models of emotion regulation^{1,26}. Most trait questionnaires capture habitual tendencies across multiple temporal stages, including encountering and recognizing reappraisable situations, deciding whether to regulate, and selecting reappraisal over competing strategies. On the one hand, previous work showed associations between reappraisal habits, ability, and neuro-behavioral implementation^{15–17}, including causal evidence that neuromodulation affects reappraisal habits¹⁸. Similarly, we found a very strong correlation between habit and ability questionnaires, pointing towards a functional relation on a conceptual level. On the other hand, standard fMRI tasks bypass much of the naturalistic reappraisal sequence: participants are instructed when and how to regulate, receive practice trials and respond to standardized stimuli optimized for reliable group-level effects. Therefore, individuals with higher

reappraisal capability might derive their real-world advantage less from universally superior neural implementation than from selecting, initiating and sustaining reappraisal in personally meaningful contexts, and from encountering environments that more readily afford successful reappraisal³⁶. This fits broader arguments that self-regulation is situated rather than governed by a context-general regulatory resource³⁷. Future paradigms should therefore additionally engage other regulatory stages including strategy selection, personally relevant stimuli and spontaneous deployment.”

“Between persons, however, brain-behavior associations are systematically much smaller²⁰ — a dissociation also observed for non-fMRI physiological markers of emotion²⁷. Neural degeneracy provides one principled account: multiple neural configurations can support similar regulatory outcomes, and the specific configurations recruited may differ markedly across individuals³⁸, particularly in higher-order association cortices where anatomical variability is substantial³⁹. Yet degeneracy alone cannot explain the magnitude of the within-between discrepancy. Process-oriented fMRI studies routinely identify replicable regional response patterns, as reflected in meta-analyses of reappraisal², and population-level models trained to decode within-person affective states generalize well to unseen healthy and clinical cohorts⁴⁰. This indicates substantial representational overlap between individuals on a process-level. Three additional factors are likely important. First, we demonstrated that established multivariate neuro-affective signatures can produce large effect sizes while conflating emotion generation and regulation, limiting their psychological interpretability and showing no association with reappraisal capability. Second, standard reappraisal tasks were designed for process-based research and may not elicit the stable individual differences needed for strong brain-behavior associations²⁴. In support of this, we found little variance on the between-person level in task-based affective ratings, even in two large samples with population representative variance on affective traits²⁰. Third, we found that between-person fMRI responses are strongly shaped by person-level whole-brain responsivity — potentially owing to factors such as neurovascular variation²¹ — producing highly correlated regional responses that limit anatomical specificity. This extends to canonical resting-state networks, where correlations reached .50 to .80 across multiple task domains. This complicates efforts to interpret individual differences in task fMRI responses through influential clinical models, including the triple-network model, suggesting antagonistic relations between canonical networks⁴¹. These global effects are not removed by calculating contrasts between conditions. As a result, most between-person variance in neural responses likely reflects confounded individual differences, in stark contrast to task-based affective responses. This highlights a fundamental misalignment, potentially alleviated by tasks designed to specifically induce individual differences while accounting for whole-brain responsivity^{20,42}.”

R1-C2:

Relatedly, the authors argue that the lack of an association of the three measures is due to them being interindividual associations and the methods not possessing enough interindividual variability partly because they were not constructed for assessing that variability. The fantastic chance that the authors have is to actually test that interpretation for two of the measures, brain activity and task-related affect ratings, as these were acquired simultaneously across the duration of the experiment. So intra- and interindividual variation is available and could be compared. We understand that this would require

reanalysis of the first level fMRI data together with the ratings but believe it would be an important step.

Response:

We thank the reviewers for stimulating these insightful follow-up analyses! We performed such a decomposition for task-based affective ratings and the two main neural targets (amygdala and regulatory network) in two large samples, including over 350 individuals and 16,000 trials. This was a pragmatic decision, as single-trial analyses were not feasible with the full consortium, while this sample size grants good precision for these decomposition analyses. Notably, we previously found that these two community samples have population representative variance in negative affective traits, making them particularly well-suited in our opinion. The results clearly supported the hypothesis. The corresponding addition to the results section is shown below:

“The profound individual differences in global signal responses are in conflict with the argument that most fMRI tasks have been designed to elicit strong within-person effects, at the cost of meaningful between-person variance, despite their common use in individual difference research²⁴. We tested this potential misalignment by analyzing trial-wise responses to negative pictures in two large samples with population representative variance on negative emotional dispositions, performing the most common reappraisal task²⁰. For trial-wise task-based affective ratings, the majority of variance was at the within-person level (92-93%; Figure 4). In contrast, for neural responses in the amygdala and regulatory networks, about 60% of variance was on the between-person level. Hence, associations are likely reduced because common fMRI task designs prioritize robust within-person condition effects over between-person effects²⁴, a mismatch further amplified by large individual differences in global signal responses.”

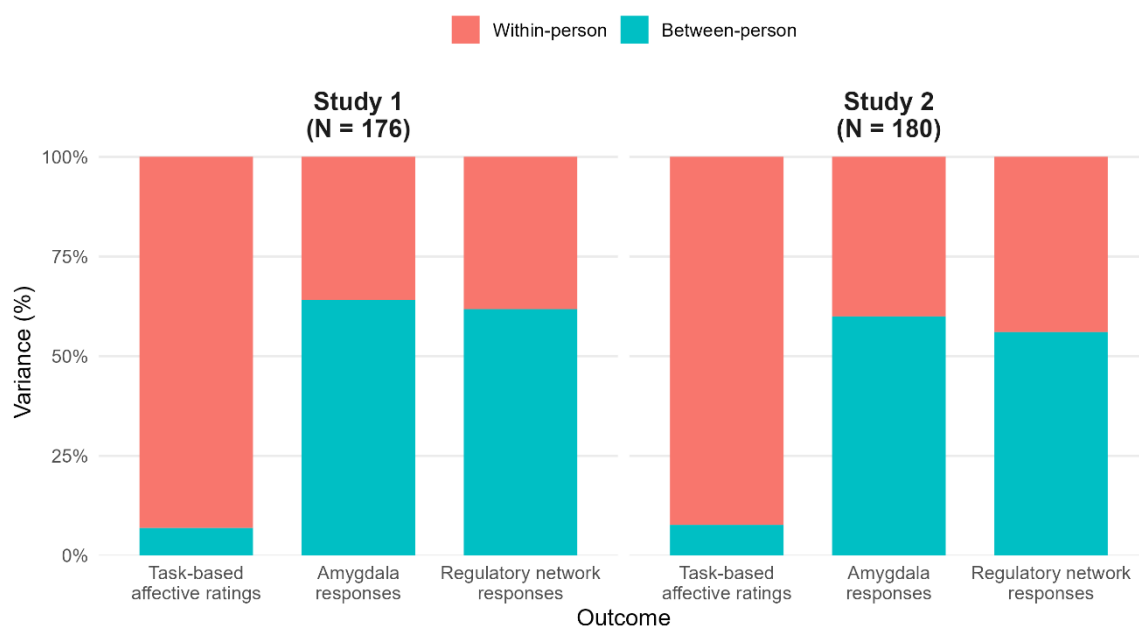


Figure 4. Variance within versus between participants.

Variance decomposition for three focal outcomes based on two large studies (Table S1, Gianaros et al. 2020, AHAB-II/PIP). Single trial responses were extracted for the three outcomes. The

variance in trial-wise responses was decomposed using an intercept-only linear mixed model, with a random intercept term reflecting variance between participants. To allow stable estimates, for each participant, trial responses outside three times the interquartile range were removed. Then, whole participants with average responses outside three times the interquartile range were removed.

R1-C3:

The authors discuss in detail the possibility of overall interindividual differences in brain activity „overshadowing“ any reappraisal related differences and thereby preventing significant brain-behavioral correlations. What did not become clear to us is intraindividual contrast images (e.g. reappraisal vs. view) do not control for this? Does the data indicate an interaction of overall activation differences with experimental condition?

Response:

The reviewers raise an important question that deserves explicit clarification, as we personally find the answer is not immediately intuitive.

Intraindividual contrast images of the form [regulate – view] do subtract the mean response in the view condition from the mean response in the regulate condition *within each person*. This removes within-person differences between conditions — which is precisely what makes them useful for detecting group-level task effects. However, they do not remove between-person differences in overall BOLD responsivity.

To see why, consider two individuals who differ in their global neurovascular coupling — for example, due to differences in age, cardiovascular fitness, or hematocrit levels (Krishnamurthy et al., 2020; Ward et al., 2020; references included in the manuscript; Krishnamurthy et al., was added, as it demonstrates the phenomenon particularly well for contrasts between two conditions). Both individuals A and B have identical differences between contrasts in their true underlying neural responses (Condition 1 = 1, Condition 2 = 2). Individual A generates BOLD responses that are, on average, 30% larger than individual B's across all conditions and regions. When the contrast [regulate – view] is computed, both the regulate and view responses are scaled by this person-specific factor. The contrast response between both conditions of individual A would be 1.3, the contrast response of individual B would be 1. The subtraction removes the *additive* baseline difference, but the *multiplicative* scaling — the global responsivity factor — remains embedded in the variance of the contrast image itself. A person with uniformly stronger BOLD responses will tend to show stronger contrast values in both directions, regardless of the specific cognitive or emotional processes of interest.

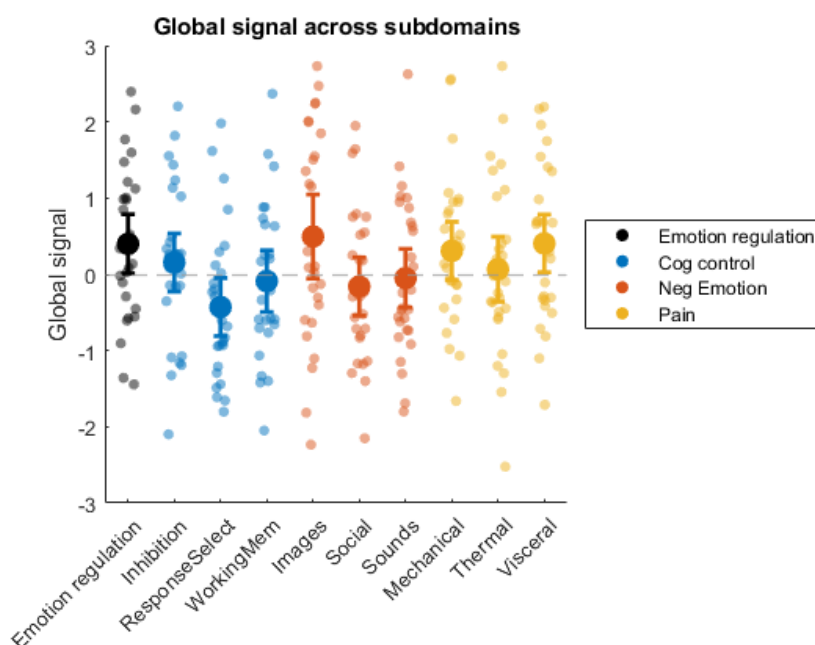
We agree with the reviewers that such effects might lead to an interaction between overall activations and experimental conditions, for example, if neurovascular factors and experiment-related psychological processes are confounded (e.g., if age influences cerebral blood flow and cognitive capacities relevant for a task). We tested this for the multimodal datasets already used in the previous submission and found no clear systematic advantage of “higher load” conditions versus control conditions (e.g., regulation, exerting cognitive control, negative emotional stimuli, pain). The results are shown in the figure below. Overall, we believe dedicated methodological studies will be important to further elucidate such effects.

The Results section now makes this concrete:

"Contrary to theoretical predictions, participants with stronger amygdala responses during reappraisal also showed substantially stronger responses throughout the rest of the brain (Figure S6; mean $r=.27$, range = $-.08$ to $.68$). This included both regulatory networks (first: $r=.29$, $p<.001$; second: $r=.24$, $p<.001$; $k=35$, $N=2088$) and over 170,000 voxels brain-wide, with only a single voxel showing a negative correlation (Figure 2D). A person with a larger amygdala response than others therefore also tends to show larger responses in the dlPFC and elsewhere — making the expected opposing pattern almost undetectable with standard fMRI individual-difference analyses."

In the Discussion, we further note:

"These global effects are not removed by calculating contrasts between conditions."



Note. Y-axis shows person-wise global signal responsivity, scaled on the study-wise standard deviation to bring responses to a comparable scale, regardless of scanning parameters. After separate scaling, two studies from the same subdomain were combined for plotting to facilitate clearer visualization. Error bars represent 95% confidence intervals of the mean.

R1-C4:

Minor:

L. 128, first sentence: error in Referencing

Response:

Corrected.

L. 138f. Reappraise vs. permit emotion – is “permit emotion” the “watch” condition?

Response:

Yes, we tried to clarify by writing “view-without-regulation”.

L. 238 Jackknife procedure: Why do both studies have a large impact? Is it because of their sample size? Maybe turn the paragraph around and say it in the beginning. This way the reader knows in the beginning why they are influential

Response:

We restructured the paragraph and noted that sample size but also study-specific design effects likely lead to its influence. If its influence were only due to sample size, the expectation would be to observe fewer significant voxels after its exclusion. In contrast, more significant effects were observed after its exclusion, indicating design-specific effect suppression by this study in the global analysis, due to its large sample.

L. 246f. Is the neurosynth analysis presented anywhere in the paper?

Response:

We thank the reviewers for pointing out this shortcoming and provide more comprehensive information on these analyses:

“To facilitate the psychological interpretation of the significant neural activation pattern, we conducted automated meta-analytic decoding using Neurosynth, which quantifies the association between whole-brain activation maps and large-scale term-based meta-analytic patterns derived from the neuroimaging literature (e.g., “pain”, “fear”, “motoric”). Neurosynth decoding against 50 topic maps from 11,406 fMRI studies indicated the meta-analytic activation map is most likely related to language processing.”

L. 281 reappraise vs. view – maybe stick with this term instead of “permit emotion” which sounds unfamiliar to me and inhibits reading flow

Response:

As noted above, we changed both instances of “permit” to “view-without-regulation”.

**L. 281f. “only negative signs” – maybe one sentence about the meaning here?
Response:**

We added the following sentence:

“Therefore, brighter colours correspond to larger negative associations between neural responses and indices of reappraisal capability.”

L. 289 bracket missing

Correlations have been indicated with two numbers after the decimal point – maybe this is due to the journal’s statistic reporting guide?

Response:

The bracket was added. Choosing two decimal points for correlations was based on APA 7. We personally find this strikes a good balance between readability and sufficient precision. Specifically for p-values, the manual suggests 2-3 decimal points. Here, we chose the latter version with 3 decimal points, as we believe a more precise appraisal of values between .05-.001 is important.

REVIEWER 3

R3-C1:

This study addresses the construct validity of “reappraisal capability” by examining the interrelationships between three hypothesized ways of measuring it, including trait questionnaires of habitual reappraisal, reduced self-reported negative affect when presented with negative stimuli in tasks, and reduced amygdala activity when presented with negative stimuli in tasks.

The immense amount of collaboration required to do this project is laudable. All teams used a uniform set of procedures to calculate intercorrelations between trait, self-report, and neural measures. For the analytical approach, the authors calculated correlations between measures across individuals for each study separately. They then used these statistics in subsequent analyses. This approach is distinct from more recent “mega-analysis” approaches where a single model is used to combine data across sites and studies at the subject level, but it is reasonable given the heterogeneity of the datasets and measures. Regardless, the volume of data and the approach is certainly sufficient to test with confidence whether there is support for even a small correlation at the population level (see point 7 below for additional support). Additional analyses examined other brain regions and also correlations between brain regions in this data and also in another multi-task dataset, that more generally challenge the assumptions of the conventional “up/down” regulation model that underlies the neuroscience of emotion regulation.

This is an exceptional paper that will undoubtedly generate a great deal of discussion, debate, and impact in the field. It is well-executed and contributes field-challenging conclusions that researchers will have to reckon with. The manuscript is clearly and concisely written. I only have minor comments for the authors below.

Response:

We thank Reviewer 3 for the careful and generous reading of the manuscript. We note that Reviewer 3 specifically highlighted two aspects of our approach that have been questioned elsewhere in this revision: first, that the meta-analytic strategy — while distinct from mega-analysis — is “reasonable given the heterogeneity of the datasets and measures”; and second, that the additional analyses examining between-person correlations across brain regions “more generally challenge the assumptions of the conventional ‘up/down’ regulation model.” We agree with this characterization and have strengthened both of these contributions in the revised manuscript in response to comments from all reviewers, including custom simulations on mega-vs meta-analytic approaches, reported in the supplements.

Minor:

R3-C2:

Given the lack of convergence between measures, the authors' main suggestion for future work is that the field needs better alignment between neural measures and constructs. As they note in the discussion section: "The weak convergence between common outcome measures and their inconsistent associations with canonical reappraisal networks underscore the need for better-aligned behavioral, trait-based, and neural constructs for them to capture a shared core of cognitive reappraisal." Here, the authors could also consider that the theories and statistical approaches that underlie this work assume relatively simple mappings: e.g., dominantly linear or at least monotonic functions, and one-to-one mappings between brain regions/networks and psychological phenomena. Their methodological approach makes these same assumptions, even if it is to show that this approach ultimately leads to challenges in construct validity. But there is also an alternative future direction. Increasing work in affective neuroscience suggests that the mind–brain mapping might be more complex (i.e., may involve degeneracy or multiple solutions). By this account, refinement of measures may still not lead to clear relationships. Rather, alternative theoretical and analytical approaches are required that account for complex mappings. Note, this is not a criticism of the authors' work—but rather a comment that could be discussed in the review process only or could be considered by the authors in a revision to stated future directions.

Response:

We thank the reviewer for raising this important theoretical point, which we have incorporated into the revised Discussion as an additional explanatory framework for the observed null findings. While we pointed towards the need and options to address degeneracy in future studies, we also believe the discrepancy between very successful within-person decoding and much smaller effect sizes in between-person analyses raises further question beyond (not instead of) degeneracy, which we tried to balance in the added sections:

"Between persons, however, brain-behavior associations are systematically much smaller²⁰ — a dissociation also observed for non-fMRI physiological markers of emotion²⁷. Neural degeneracy provides one principled account: multiple neural configurations can support similar regulatory outcomes, and the specific configurations recruited may differ markedly across individuals³⁸, particularly in higher-order association cortices where anatomical variability is substantial³⁹."

We also note an important counterargument to the idea that degeneracy alone fully accounts for the null findings, namely, that within-person decoding succeeds in generalizing to new unseen individuals despite degeneracy:

"Yet degeneracy alone cannot explain the magnitude of the within-between discrepancy. Process-oriented fMRI studies routinely identify replicable regional response patterns, as reflected in meta-analyses of reappraisal², and population-level models trained to decode within-person affective states generalize well to unseen healthy and clinical cohorts⁴⁰. This indicates substantial representational overlap between individuals on a process-level."

R3-C3:

I think the results section could use a bit more clarification on the measures they are using for statistical analysis. Information from this sentence from the methods section would be helpful to reiterate or paraphrase in the first subsection of the results section: “Research groups prepared up to three unthresholded group-level statistical images, containing voxel-wise t-statistics for the between-person correlation of the whole-brain neural response in the [reappraise - view] contrast and the three outcomes of interest.”

Response:

We thank the reviewer for this observation. We carefully revised the Results section by integrating more relevant aspects of the Methods section in order to provide more rationale and clarification.

The revised Results section now opens with the following paragraph:

“In a first step, we aimed to assess the overlap between the three main indices of cognitive reappraisal capabilities: trait questionnaires, task-based affective ratings following reappraisal, and amygdala down-regulation. If these three indices represent a similar core construct, or are otherwise functionally related, they should show a meaningful association. Individual consortium study teams calculated the correlations between the three indices of reappraisal capabilities based on a standardized coding protocol using functionally predefined regions-of-interest, ensuring that higher values on each index indicate higher reappraisal capabilities. Study-level results were aggregated with random effects meta-analyses, which is statistically equivalent to person-level random effects mega-analysis for the preregistered analyses and was intended to circumvent data sharing thresholds (see Supplementary Methods section: “Meta- vs mega-analytic approaches”).”

R3-C4:

“A person’s ability to either down-regulate their amygdala or up-regulate regulatory regions is often interpreted as an indicator of interindividual reappraisal capability” — consider rephrasing slightly since the current phrasing can be read in two different ways, one of which implies dualism.

Response:

We changed it to:

“Theoretical models predict that effective reappraisal should be associated with lower amygdala and higher prefrontal responses — an opposing pattern. Between-person differences in overall brain responsivity^{20,21}, however, could blur such regional specificity.”

R3-C5:

This part of the results section could use a bit more clarification: “...people differ in their global brain responses, likely due to biological confounders, which might dominate any specific regional effects and dilute their usefulness as a marker of an individual’s emotion

regulation capability.” In general, the “Amygdala down-regulation” section could be developed further to clarify the argument and the counterargument.

Response:

We made major changes at the beginning of the section (see below) and minor changes in the paragraphs that follow on from this section in the manuscript (not shown here):

“Theoretical models predict that effective reappraisal should be associated with lower amygdala and higher prefrontal responses — an opposing pattern. Between-person differences in overall brain responsivity^{20,21}, however, could blur such regional specificity.

Contrary to theoretical predictions, participants with stronger amygdala responses during reappraisal also showed substantially stronger responses throughout the rest of the brain (Figure S6; mean $r=.27$, range = $-.08$ to $.68$). This included both regulatory networks (first: $r=.29$, $p<.001$; second: $r=.24$, $p<.001$; $k=35$, $N=2088$) and over 170,000 voxels brain-wide, with only a single voxel showing a negative correlation (Figure 2D). A person with a larger amygdala response than others therefore also tends to show larger responses in the dlPFC and elsewhere — making the expected opposing pattern almost undetectable with standard fMRI individual-difference analyses.”

R3-C6

Is the positive correlation true across tasks for default mode and frontoparietal networks, too?

Response:

We thank the reviewer for pointing us towards these analyses! The correlations are in fact even larger for resting state networks. We believe the point raised by the reviewer is particularly important, as it also relates to influential triple network theories of psychopathology. The correlations suggest that, at least for task-based responses, these networks lack discriminant validity. Therefore, we added these analyses to the revised manuscript and made the following additions to the results section:

“Crucially, this was not specific to reappraisal: the same pattern held across 18 studies spanning negative emotion, pain, and cognitive control⁹ (mean $r = .37$; Figure 3A), with between-person correlations among canonical resting-state network responses²³ reaching $.50-.80$ (Figure 3B).”

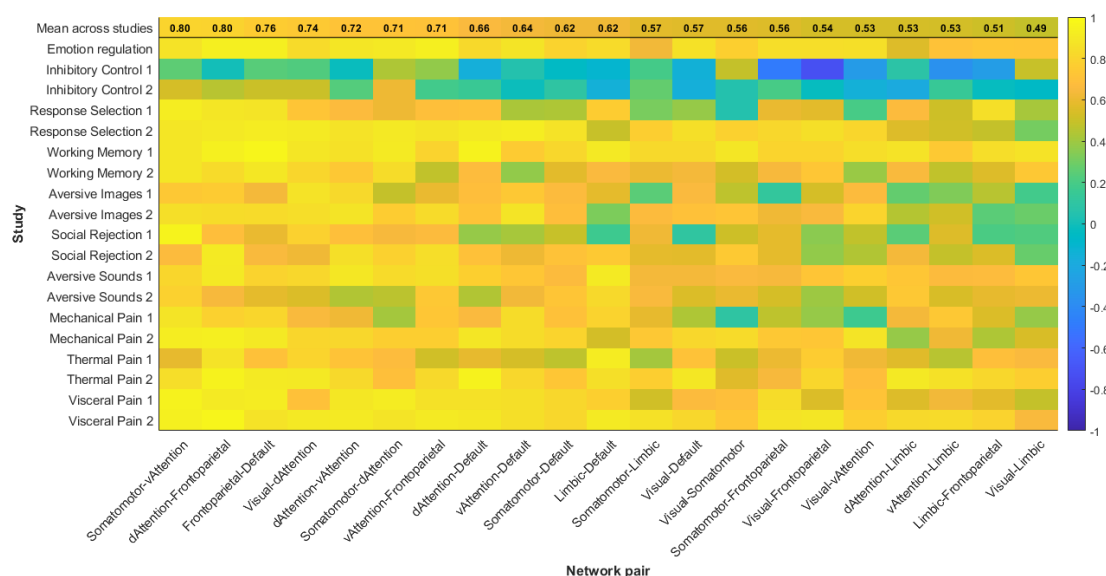


Figure 3B. Between-person correlations for responses in pairs of resting-state networks (x-axis) in the same multimodal dataset. Resting-state networks are based on the parcellation by Yeo and colleagues²³. The top row indicates the average correlation across studies. d/vAttention = dorsal/ventral attention networks. Default = Default mode network.

We further added the following paragraph to the revised Discussion:

" Third, we found that between-person fMRI responses are strongly shaped by person-level whole-brain responsivity — potentially owing to factors such as neurovascular variation²¹ — producing highly correlated regional responses that limit anatomical specificity. This extends to canonical resting-state networks, where correlations reached .50 to .80 across multiple task domains. This complicates efforts to interpret individual differences in task fMRI responses through influential clinical models, including the triple-network model, suggesting antagonistic relations between canonical networks⁴¹."

R3-C7:

The “statistical power” section seems a bit detailed for the main text and could be moved to supplementary, and the main text revised to simply state that numerous approaches were performed to validate statistical power.

Response:

We removed technical details of power analyses from the main text to the supplements and now only state the results in the former. Notably, there were important changes to the power analysis based on comment R3-C8.

R3-C8:

To further convince skeptical specialist readers that they have enough data, one additional and relatively simple analysis they could include is to generate data simulations assuming

the 40 studies with their specific sample sizes and a true population correlation of .1 or so, and see if their simulated data rejects a null hypothesis. While I was convinced by the authors' thoroughness already, I did this in addition and found it useful.

Response:

We thank the reviewer for this comment. Performing these analyses showed an imprecision of our previous approach, which we now replaced with an approach aligned with that suggested by the reviewer.

Specifically, our previous approach provided power-analysis based on Bonferroni-Holm FWE correction, rather than Benjamini-Hochberg FDR correction, which was preregistered, used throughout the manuscript, and affords higher statistical power.

Below, we present the additions explaining the present power simulation approach.

Results:

Power in voxel-wise analyses was mostly >80% in simulated scenarios for this effect size, especially for the network of interest approach (see Supplementary Methods section: "Power Analysis")."

Supplements:

We performed simulation-based sensitivity analyses to assess the statistical power of our voxel-wise analyses to detect a small population correlation of $r=.10$. We simulated scenarios with an increasing number of true clusters (1–4). Between-study heterogeneity was set to $\tau=.06$, corresponding to the third quartile in our empirical data, and supplemented by an additional simulation at $\tau=.08$ as a more conservative scenario.

Four anatomically distinct, non-contiguous clusters were defined within the preregistered network of interest. Cluster sizes were 800, 500, 300, and 200 voxels. To avoid artificially inflated power estimates that would arise from assuming independence across clusters, study-level cluster effects were simulated jointly according to the empirical correlation matrix of these four regions.

Power analyses were conducted for four nested scenarios: first, only the largest cluster was assigned a true effect; second, the two largest clusters; third, the three largest clusters; and fourth, all four clusters. Each scenario was simulated across 1,000 iterations. Power was quantified as the proportion of iterations in which at least one true cluster was detected after false discovery rate correction, either within the preregistered network of interest or across the whole brain. The results are shown below in Figure S11.

While the reliable ground truth of typical between-person effects in fMRI is unclear (Cremers et al., 2017), our cluster sizes were guided by those reported in previous fMRI meta-analyses on group differences in emotion regulation (Morawetz et al., 2025; Picó-Pérez et al., 2017)."

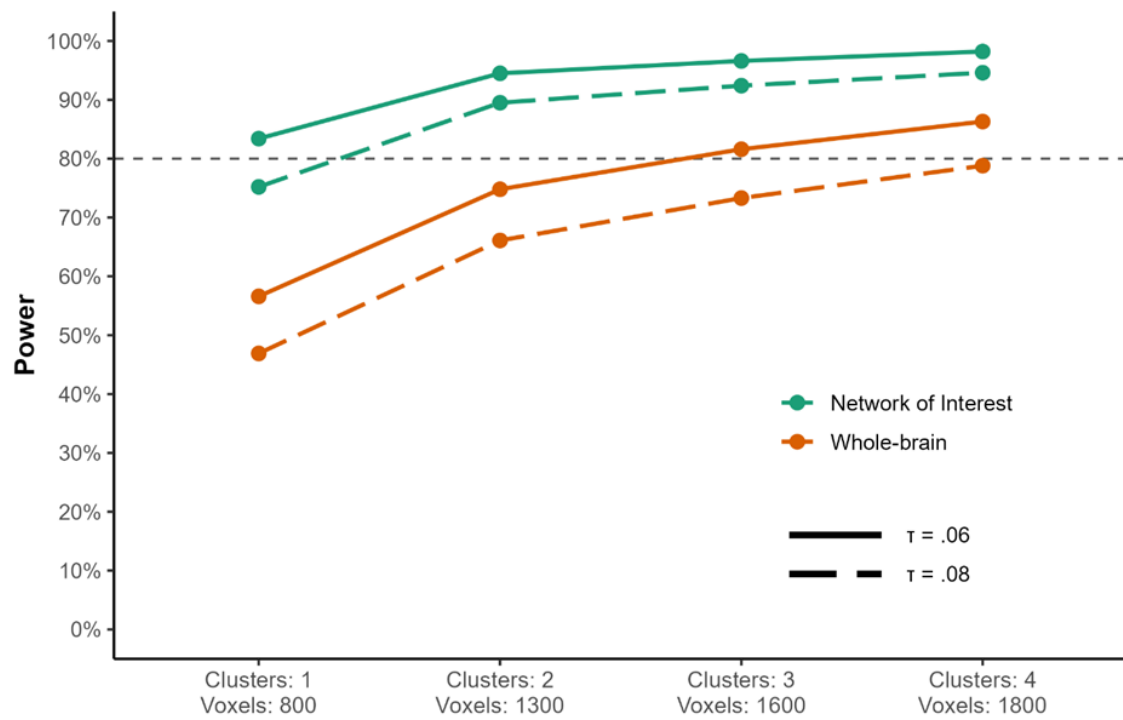


Figure S11. Statistical power to detect at least one true cluster as a function of the number of simulated clusters (and corresponding total voxel extent). Power is shown for FDR correction within the preregistered network of interest and across the whole brain. τ denotes between-study heterogeneity.

REVIEWER 4

Overall comments

This paper combines data from over 40 different fMRI Studies of reappraisal to address questions about the extent to which behavioral and brain measure of reappraisal correlate with one another. Given that emotion regulation is a core ability central to healthy adaptive functioning, and that problems with regulation characterize nearly every common clinical disorder, this paper addresses an important topic.

That said, I do not believe this paper in its current form is ‘complete’, methodologically and conceptually thorough and suitable for publication. The questions the paper seeks to address - about the nature and correspondence of behavioral and neural measures of reappraisal as a paradigmatic emotion regulation strategy - are important *and* very complex. As such, addressing them appropriately demands a great deal of context and nuance - both analytically and - perhaps as critically - conceptually, when discussing the core psychological constructs and neural systems of interest.

The lack of nuance/consideration of complex analytic and conceptual issues muddles the paper’s message and potential impact. The paper reports largely null results, mentions potential reasons for this such as commonly used reappraisal tasks not being ecologically valid, and notes that there may also be widespread issues in fMRI research like multiple regions showing high degrees of positive correlation (i.e. inter-regional comparisons are confounded by so-called global signal issues). The paper does not, however, unpack what any of this means. This, I believe, is absolutely crucial. A paper of this kind must, I believe, consider the ways in which:

(Point a) the ways in which the current analytic approach - which relies entirely on study-level correlation maps - might be limited in terms of addressing the core issues.

(Point b) the conceptual complexity of the core issues the paper purports to tackle - ranting from a theoretical analysis of why one would or would not expect different behavioral and neural measures of reappraisal to correlate to what it means to talk about emotion regulation capability.

(Point c) the design / approach of current reappraisal studies may or may not contribute to the apparent null findings (e.g. is this an issue of approach / method - could future studies could be different, e.g. more ‘ecologically’ valid?). As opposed, that is., to the point b), where the issue is that different behavioral and neural ways of assessing regulation really would not be expected to relate to one another, based on a first principles analyses of what different measures reflect?

Detailed comments

(Point a)

With respect to core questions about the current analytic approach:

R4-C1:

— A key question is whether the current analytic approach is the appropriate one for addressing the core questions at hand about brain correlates or reappraisal corresponding to behavioral measures of reappraisal. The present approach states that it takes account of an N of over 2000 going into the main analyses, which involve the aggregation of study-level correlation maps for 41 fMRI Studies of reappraisal. Given that study level maps - not individual subject data - are entered into analyses - it does not seem appropriate to state that these analyses were based on an end of over 2000 participants. The real N is, in essence, the number of studies included in the paper, which is 41.

That said, the core conundrum is whether results would be different if the authors did their analyses differently.

First and foremost is the question of what the results might be if the authors focused not on study-level maps, but had access to the actual subject-level imaging data. This is important for many reasons. Perhaps chief among them is the possibility that truly meaningful individual differences in underlying neural dynamics might only be observable in subject-level data - in the ways that each individual relies on different combinations of key underlying neural systems when reappraising. By relying on study level maps, the current analysis is insensitive to such within individual differences. As such, the key sources of variability in the neural systems underlying reappraisal may be obscured by the averaging that goes into study-level maps.

Response:

We thank the reviewer for raising this fundamental methodological question, which we highly appreciate. We address these concerns on three levels: (1) the conditions of practical statistical equivalence of our meta-analytical approach with mega-analysis (i.e., the reviewer's suggested analysis of subject-level imaging data), (2) the reason for choosing meta-analysis, (3) and the separate question of neural heterogeneity across individuals.

(1) The statistical equivalence of our approach with mega-analysis

On the meta- vs. mega-analysis equivalence: for voxel-outcome correlation analyses, image-based meta-analysis and subject-level mega-analysis are approximately statistically equivalent under standard random-effects assumptions. This equivalence follows from the mathematical structure of both models and is confirmed empirically by prior simulation work (Koi et al., 2021; Eisenhauer et al., 2021), which we also noticed within our own unpublished sensitivity analyses of mega-analysis on neural signatures including the PINES, motivating the present strategy (Sicorello et al., 2021). We have now added explicit simulations demonstrating this equivalence for our specific setting (added to Supplementary Methods). These simulations further confirm the practical equivalence of these two approaches for voxel-outcome correlations at our sample sizes.

Notably, as supported by their relative equivalence, the total number of participants (N) is an equally crucial determinant of statistical power for both meta- and mega-analytic approaches under random effects assumptions. The number of studies (k) is also similarly relevant for both approaches if between-study heterogeneity is present, which was relatively low across analyses.

Therefore, we made sure in the revision that both N and k are mentioned together throughout the manuscript. This is in agreement with the PRISMA guidelines (Page et al., 2021; e.g., see abstract checklist).

In the results section, we added:

“Study-level results were aggregated with random effects meta-analyses, which is statistically equivalent to person-level random effects mega-analysis for the preregistered analyses and was intended to circumvent data sharing thresholds (see Supplementary Methods section: “Meta- vs mega-analytic approaches”).”

The following additions were made to the supplements in a dedicated section:

“We performed custom simulations to demonstrate that the relative equivalence of random effects mega- and meta-analysis holds for our targeted analyses. Datasets were simulated with a population correlation of $r = .20$ and between-study heterogeneity of $\tau = .10$. The number of simulated studies and their respective sample sizes were fixed to those for the voxel-wise correlations with trait questionnaires ($k = 37$, $N = 2091$). In total, 2000 datasets were simulated.

Mega-analysis was performed using lme4 (Bates et al., 2015). Study-wise data were z-transformed using their individual mean and standard deviation. This brings mega-analytic estimates to the same standardized scale as the meta-analytic estimates to allow clear comparisons. Moreover, this is a reasonable preprocessing step in fMRI mega-analysis due to arbitrary scale differences between studies given different fMRI measurement protocols (Sicorello et al., 2021). It further removes the necessity for fixed and random intercepts, as they are set to 0 for individual studies by z-transformation, removing the bias that mega-analysis would otherwise have to estimate a larger number of parameters than meta-analysis. The correlation-metric slopes were allowed to vary across studies (i.e., random slopes), which mirrors the assumption of between-study heterogeneity in effect sizes of random effects meta-analysis, which is a crucial step for both approaches (Barr et al., 2013; Costafreda, 2009).

The main results are shown below (Figure S12). The left panel shows the mean correlation estimate and its standard deviation across simulations. The discrepancy between mega- and meta-analytic estimates was minimal at $\Delta r = .006$. Both approximated the population effect size well. The right panel shows the precision of the estimates, operationalized as the width of 95% confidence intervals. The difference was again minimal at $\Delta 95\% \text{ CI} = 0.003$.”

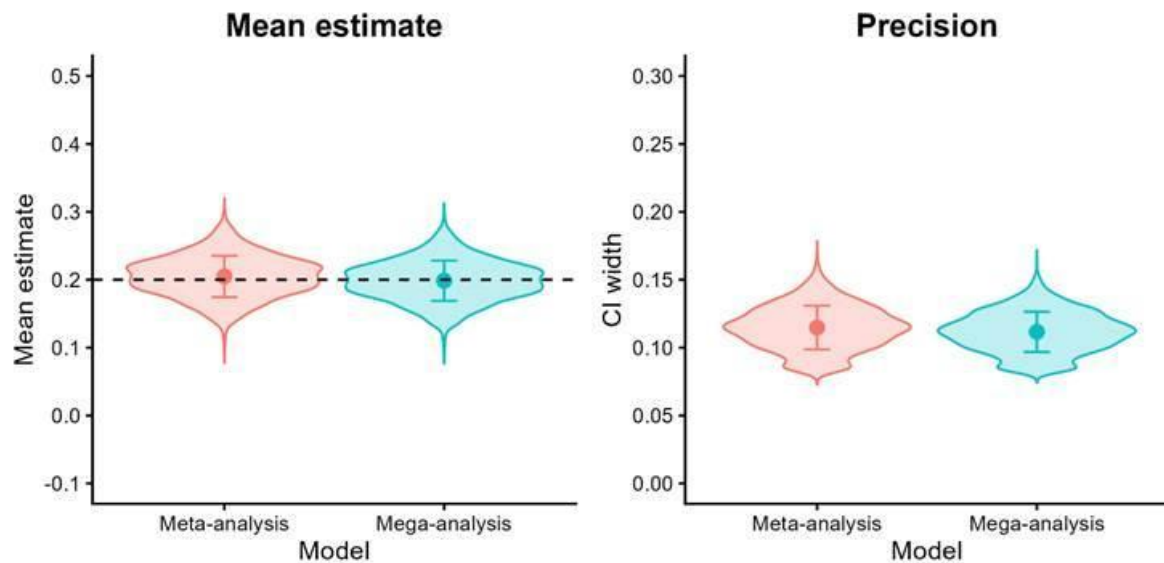


Figure S12. Mean estimates and precision comparison between meta- and mega-analysis across 2000 simulated datasets.

(2) The potential practical advantages of meta-analysis

We actively chose the meta-analytic approach because it likely allowed us to gather more data than a mega-analysis would have. A consortium survey confirmed substantial hesitation for sharing fMRI images, despite social desirability and selection biases in the survey (Paret et al., 2022). In our experience and based on specific discussions with potential consortium members at the planning stage, this hesitation extended to subject-level contrast images due to varying institutional IRB interpretations of data protection. Therefore, our reasoning was that if mega- and meta-analysis are approximately equivalent for our target analyses, we can gather more data, achieving higher precision and statistical power, by only requesting study-level images, if this increases the readiness of sharing data. We added this reasoning and simulation-based power for different scenarios to the supplements.

“Figure S13 further shows how statistical power would have differed if a lower number of studies was included due to data sharing barriers, exemplary for a population correlation of $r = .10$ across 1000 true positive voxels. For simplicity, each included study had a study-wise sample size of $n = 57$, corresponding to the average sample size in our empirical consortium sample. A lower number of studies would have substantially reduced statistical power, with increasing between-study heterogeneity.”

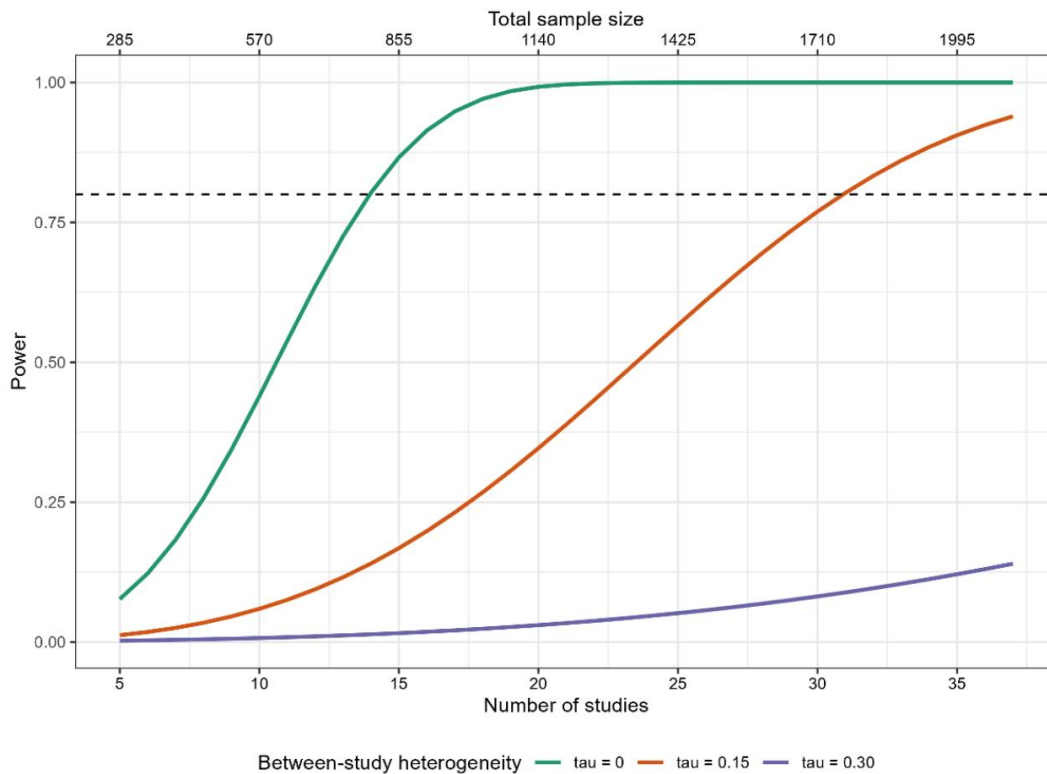


Figure S13. Statistical power as a function of the number of included studies (bottom x-axis) with corresponding total sample sizes (top x-axis; assuming $n = 57$ per study). Results are shown for three levels of between-study heterogeneity τ and a fixed population correlation of $r = .10$. $\tau = 0$ (green) indicates a perfectly homogeneous scenario (all studies share the same true effect), while $\tau = 0.15$ (orange) and $\tau = 0.30$ (purple) reflect increasing variability in true underlying effects across studies. The dashed horizontal line marks the conventional 80% power threshold.

(3) Neural heterogeneity

The reviewer's observation that individuals may rely on different neural systems during reappraisal is theoretically important. This concern relates closely to neural degeneracy, which we address in detail in our response to R3-C2 and now discuss explicitly in the revised manuscript. We agree that such heterogeneity could attenuate associations in voxel-wise analyses, particularly if different individuals achieve similar regulatory outcomes through partially distinct neural configurations.

At the same time, we believe this point raises a distinction between two related but separable questions. Characterizing individual-specific neural topographies, subgroup structure, or distinct neural routes to successful reappraisal would indeed require subject-level maps and person-centered approaches such as clustering or decomposition. Our primary question, however, was whether interindividual variation in neural engagement at a given region (or voxel) covaries with interindividual variation in reappraisal-related behavioral measures. This voxel-wise covariation requires only that there is some spatially consistent component of neural engagement that covaries with the behavioral measure, relating to the larger fMRI research field relying on the assumption of partially generalizable neural representations in specific anatomical modules.

We therefore acknowledge complete spatial non-overlap across individuals as a limiting case in which voxel-wise covariation analyses could miss meaningful individual differences. However, we consider this limiting case unlikely to fully explain the present findings. More robust findings in process-oriented fMRI meta-analyses indicate that reappraisal and affective responding show a degree of representational similarity across individuals. This is also supported by the cross-individual generalization of population-level multivariate signatures, which decode affective states in held-out healthy and clinical samples. Thus, while neural degeneracy may contribute to the observed null findings, it is unlikely to fully explain the magnitude of the within-between discrepancy.

“Between persons, however, brain-behavior associations are systematically much smaller²⁰ — a dissociation also observed for non-fMRI physiological markers of emotion²⁷. Neural degeneracy provides one principled account: multiple neural configurations can support similar regulatory outcomes, and the specific configurations recruited may differ markedly across individuals³⁸, particularly in higher-order association cortices where anatomical variability is substantial³⁹. Yet degeneracy alone cannot explain the magnitude of the within-between discrepancy. Process-oriented fMRI studies routinely identify replicable regional response patterns, as reflected in meta-analyses of reappraisal², and population-level models trained to decode within-person affective states generalize well to unseen healthy and clinical cohorts⁴⁰. This indicates substantial representational overlap between individuals on a process-level. Three additional factors are likely important.”

Eisenhauer, J. G. (2021) Meta-analysis and mega-analysis: A simple introduction. Teach. Stat. 43, 21–27.

Koile, E. & Cristia, A. (2021) Toward Cumulative Cognitive Science: A Comparison of Meta-Analysis, Mega-Analysis, and Hybrid Approaches. Open Mind 5, 154–173.

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ; 372:n71

Paret, C., Unverhau, N., Feingold, F., Poldrack, R. A., Stirner, M., Schmah, C., & Sicorello, M. (2022). Survey on open science practices in functional neuroimaging. NeuroImage, 257, 119306.

Sicorello, M., Herzog, J., Wager, T. D., Ende, G., Müller-Engelmann, M., Herpertz, S. C., ... & Niedtfeld, I. (2021). Affective neural signatures do not distinguish women with emotion dysregulation from healthy controls: A mega-analysis across three task-based fMRI studies. Neuroimage: Reports, 1(2), 100019.

R4-C2:

Second, there is a question of why the authors focus only on changes in amygdala response rather than multivariant patterns reflecting changes in emotional experience. As described in other comments below, the amygdala is seen by a few effective neuroscientist today to be a strong direct contributor to emotional experience. Add Andersen and Elizabeth Phelps a very nice data from over 20 years ago showing that daily emotions reported by patients who lack the amygdala are pretty much the same as those found and healthy individual individuals. More recently, multi various neural signatures like the PINES signature from Chang et al, have been validated to correspond to variations and self-reported negative

affective experience elicited by the kinds of stimuli used in reappraisal studies. You can take the amid out of those neural signature maps and it doesn't change their ability to predict experience. Did the authors consider doing an analysis with the PINES or some other neural signature?

Response:

We thank the reviewer for this important suggestion! Our choice to focus on the amygdala was motivated by the fact that it remains the most commonly used single neural outcome in individual-differences research on emotion (dys-)regulation (e.g., Morawetz & Basten, 2025; Sicorello & Schmahl, 2021), is commonly used for psychoeducation in applied contexts, and represents the most prominent target in neuromodulation approaches for emotional problems (Goldway et al., 2022; Kaleghi et al., 2025; Barksdale et al., 2025). Similar arguments were the reason to focus on prefrontal regions. In light of this ongoing focus, and especially as we shared the reviewer's skepticism concerning the amygdala (see e.g. Sicorello & Schmahl, 2021), we thought maximizing impact on the field required testing the measures that the field currently relies on, which still most prominently includes the amygdala.

We fully agree that also including machine learning based signatures is an important addition, relating to more active developments in the affective neuroscience field, in addition to the widely used canonical anatomically defined targets. We identified three types of neural signatures, which provide complementary information: (1) The PINES, as pointed out by the reviewer, as a neural measure for negative affective states, lesioned for the amygdala (expressions correlated at .99 with expressions from the full PINES). (2) A recently published signature of cognitive reappraisal, validated over multiple samples, based on comment R4-C3 (NERS; Jiang et al., 2026). (3) A recently published signature for task-based reappraisal implementation ability (i.e., individual differences; Gianaros et al., 2022). As these analyses were not feasible for all datasets, we performed these analyses on the four largest datasets. With a total sample size of N=730 and a fixed effects analysis strategy, these analyses have high precision, especially compared to correction-intensive whole-brain data and considering the high test-retest reliability of whole-brain signatures (Kragel et al., 2020). Notably, sample size was slightly smaller for some combinations, to avoid overfitting, as described in the figure legend below, but remained over N>500, still representing a comparatively large sample. Before using these signatures for individual difference prediction, we wanted to characterize how they relate to emotion generation vs emotion regulation on a process-basis, relating to the next reviewer comment (R4-C3). This revealed, in our opinion, very interesting additional findings, which we report in the main paper in a dedicated subsection. The main addition to the results section is given below:

“Finally, we tested whether multivariate neural signatures provide more sensitive indices of reappraisal capability than anatomically defined regions or networks. Neural signatures are predictive models that combine activity across many brain regions to estimate the presence or intensity of a psychological state in new individuals. We focused on three established candidate signatures: a negative affect decoder, a decoder of reappraisal as a dynamic state, and a decoder trained to predict individual differences in reappraisal success from task-based affective ratings (see Method section for details).

Using the four largest studies (N=730), we first examined whether these signatures differentiated emotion generation from reappraisal (Figure 5A, Table S3). The negative affect decoder was strongly expressed during emotion generation but increased further during reappraisal, indicating that its expression did not follow the expected reappraisal-related reduction in negative affect. The reappraisal state decoder showed stronger negative

expression during emotion generation than positive expression during reappraisal, suggesting limited specificity for reappraisal as opposed to variation in emotional state. The reappraisal ability decoder showed comparable negative expression across both contrasts. Consistent with these ambiguous task-response profiles and univariate null findings, none of the three signatures was significantly associated with reappraisal capability, whether indexed by trait questionnaires or task-based affective ratings. Descriptive correlations were near-zero ($|r| \leq .06$), with confidence intervals excluding effects larger than $|r| = .15$ and Bayes factors favoring the null hypothesis, especially for the negative affect decoder (Figure 5B, Table S4).

A similarly surprising pattern emerged for the broader systems commonly used to distinguish emotion generation from regulation. The regulatory network was comparably involved in both processes, as recently observed²⁶. Strikingly, when the outlying study 4 was removed, it even showed a much stronger preference for emotion generation ($dc = .18$, $BF_{10} > 100$) than for regulation, the latter having an undecisive Bayes factor ($dc = .08$, $BF_{10} = 0.90$). In contrast, the amygdala was selectively engaged during emotion generation and was not modulated during reappraisal ($BF_{10} = 0.03$). For detailed statistics see Table S3.”

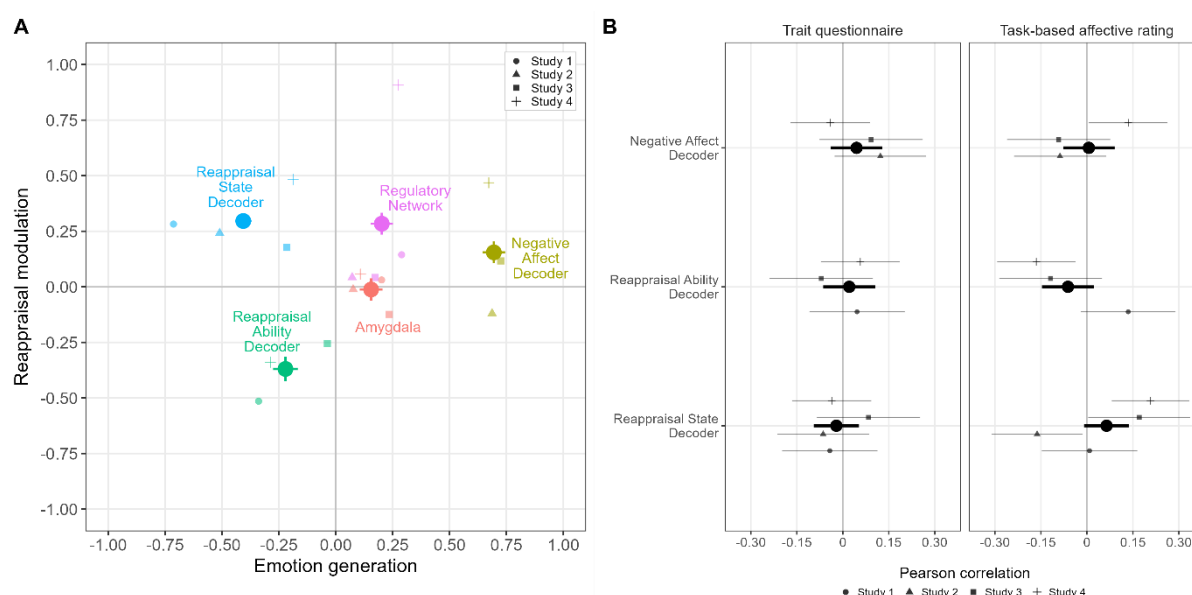


Figure 5. Neural signatures indicating reappraisal processes and individual differences.

(A) Responses of neural signatures, the amygdala, and the regulatory network to the two main contrasts: emotion generation [view negative – view neutral] and emotion regulation [reappraise negative – view negative]. Metrics in Cohen’s d . Error bars show 95% CIs. Total $N=730$, study 1 $N=180$ (Gianaros et al., 2020, AHAB-II), study 2 $N=176$ (Gianaros et al., 2020, PIP), study 3 $N=139$ (Doren et al., unpublished), study 4 $N=235$ (Brehl et al., 2021). The negative affect decoder was not used with study 1 and the reappraisal ability decoder was not used with study 2 to avoid inflated effect sizes, as these were the training samples of the previously published decoders. (B) Associations of the three neural signatures with trait questionnaires and task-based affective ratings. Sample sizes and inclusion criteria were the same as for Figure 5A.

Barksdale, B. R., Enten, L., DeMarco, A., Kline, R., Doss, M. K., Nemeroff, C. B., & Fonzo, G. A. (2025). Low-intensity transcranial focused ultrasound amygdala neuromodulation: a double-

blind sham-controlled target engagement study and unblinded single-arm clinical trial. *Molecular psychiatry*, 30(10), 4497-4511.

Gianaros, P. J., Rasero, J., DuPont, C. M., Kraynak, T. E., Gross, J. J., McRae, K., Wright, A. G. C., Verstynen, T. D., & Barinas-Mitchell, E. (2022). Multivariate Brain Activity while Viewing and Reappraising Affective Scenes Does Not Predict the Multiyear Progression of Preclinical Atherosclerosis in Otherwise Healthy Midlife Adults. *Affective Science*, 3(2), 406–424. <https://doi.org/10.1007/s42761-021-00098-y>

Goldway, N., Jalon, I., Keynan, J. N., Hellrung, L., Horstmann, A., Paret, C., & Hendler, T. (2022). Feasibility and utility of amygdala neurofeedback. *Neuroscience & Biobehavioral Reviews*, 138, 104694.

Jiang, H., He, J., Zimmermann, K., Zhou, X., Gan, X., Ferraro, S., Wang, L., Zhou, B., Li, L., Kendrick, K. M., Zhao, W., Yao, D., Yuan, T., Zhou, F., & Becker, B. (2026). Common and distinct neurofunctional signatures of dynamic naturalistic emotion regulation strategies. *Nature Communications*. <https://doi.org/10.1038/s41467-026-70708-5>

Khaleghi, A., Samiei, H., Zarafshan, H., Baloochi, S. A., & Mohammadi, M. R. (2025). Effectiveness of fMRI-based Neurofeedback Therapy on Depression: A Systematic Review. *Clinical Psychopharmacology and Neuroscience*, 23(3), 337.

Kragel, P., Han, X., Kraynak, T., Gianaros, P., & Wager, T. (2020). fMRI can be highly reliable, but it depends on what you measure (v2). *PsyArXiv*. <https://doi.org/10.31234/osf.io/9eaxk>

Morawetz, C. & Basten, U. (2024). Neural underpinnings of individual differences in emotion regulation: A systematic review. *Neurosci. Biobehav. Rev.* 162, 105727.

Sicorello, M. & Schmahl, C. (2021) Emotion dysregulation in borderline personality disorder: A fronto–limbic imbalance? *Curr. Opin. Psychol.* 37, 114–120.

Sicorello, M., Herzog, J., Wager, T. D., Ende, G., Müller-Engelmann, M., Herpertz, S. C., ... & Niedtfeld, I. (2021). Affective neural signatures do not distinguish women with emotion dysregulation from healthy controls: A mega-analysis across three task-based fMRI studies. *Neuroimage: Reports*, 1(2), 100019.

Sicorello, M. et al. (2025). The functional neurobiology of negative affective traits across regions, networks, signatures, and a machine learning multiverse. 2025.05.15.653674 Preprint at <https://doi.org/10.1101/2025.05.15.653674>.

R4-C3:

Third, the authors failed to see their purported indices of reappraisal capability corresponding with activity in any of the prefrontal systems associated with reappraisal identified in a meta-analysis. There have been recent papers publishing neural signatures associated with reappraisal. Have the authors considered, asking whether expression of these patterns might relate to changes in either self reported affect during reappraisal or in changes in the expression of neural signatures for effective responding of the sort noted in the prior paragraph?

Response:

We thank the reviewer for pointing us towards these signatures! We characterized the different signatures in reference to their involvement in emotion generation versus emotion regulation and calculated correlations to trait questionnaires and task-based affective ratings. The additions to the manuscript are reported under the previous comment (R4-C2).

(Point b)

From a conceptual perspective, the paper does not present clear hypotheses or a logic as to why one would - on the basis of theory or prior empirical results - one should or should not expect to see correspondence between the measures reappraisal included in the present paper.

R4-C4:

— The paper could take greater care in its use of terms. For example, in the introduction it is stated that, “Compounding this issue, regions activated during reappraisal also subserve other functions, making it unclear whether they reflect regulation capabilities per se...”. This comment seems to confound a description of the function of a region with the behaviors/tasks during which it is activated. A region could have a singular ‘function’ - defined in terms of the specific computational processes it implements - and may participate in multiple different behaviors. Do the authors means regions engaged during reappraisal are also engaged during other tasks, such as those involving working memory, response selection, etc? If so, it is possible that, for given region, one or more common underlying functional processes plays a key role in all of these behaviors.

Response:

The reviewer identifies an important imprecision. We have revised the passage in the Introduction:

“ Compounding this, prefrontal regions engaged during reappraisal also participate in working memory, response selection, and pain processing— limiting their specificity as emotion-related markers of regulatory function.”

R4-C5:

— The paper talks about the fact that, “...task designs may not mirror naturalistic real-life regulation, raising questions about their validity and practical relevance for important individual differences...”. This is an idea shared by many in the field, and this may be an absolutely critical issue for understanding what the present results mean, but the present paper doesn’t do much to explain why this might be the case. In what way are the tasks used in the scanner not picking up on factors that may determine emotional health and well-being in everyday life? The answer to this question - and hypotheses about it - may be essential for understanding what the present results mean.

Response:

We thank the reviewer for pointing out this important gap. We comprehensively reworked the Discussion in response to this comment as well as other conceptual comments below. The following was specifically added to address a theory-driven rationale concerning the shortcomings of common tasks:

“On the other hand, standard fMRI tasks bypass much of the naturalistic reappraisal sequence: participants are instructed when and how to regulate, receive practice trials and respond to standardized stimuli optimized for reliable group-level effects. Therefore, individuals with higher reappraisal capability might derive their real-world advantage less from universally superior neural implementation than from selecting, initiating and sustaining reappraisal in personally meaningful contexts, and from encountering environments that more readily afford successful reappraisal³⁶. This fits broader arguments that self-regulation is situated rather than governed by a context-general regulatory resource³⁷. Future paradigms should therefore additionally engage other regulatory stages including strategy selection, personally relevant stimuli and spontaneous deployment.”

R4-C6:

— **“fronto-limbic” is antiquated/outdated terminology. The concept of the limbic system is outdated. See numbers papers from Joe Ledoux (e.g. Neuron paper from ~2104) detailing why the idea of a limbic system as a functional ‘system’ is simply incorrect.**

Response:

We thank the reviewer for this important correction. We replaced “fronto-limbic” with “prefrontal circuits and the amygdala”. We have verified that no instances of 'limbic' remain in the revised text. Notably, reviewer 3 asked for between-person correlations of responses in established resting-state networks (e.g., default mode, frontoparietal). We included a figure showing these results, where the x-axis refers to a “limbic network”. This was kept to maintain consistency with the naming scheme of Yeo et al. (2011), which we adapted as a parcellation due to its continuing influence. Notably, Yeo and colleagues wrote for all network labels that these were heuristic in nature (instead of delineating functionally/cytoarchitecturally coherent modules). While we believe adhering to their nomenclature can be helpful, we are of course open to other suggestions.

R4-C7:

— **At the outset, the paper states that, “We examined three widely used indices of reappraisal capability 25: (1) self-reported habitual reappraisal (trait questionnaires), (2) task-based reductions in negative affect (task-based affective ratings), and (3) task-based reductions in amygdala activation (amygdala down-regulation). This enabled us to ask whether these indices reflect a shared underlying core construct—and whether individual differences in reappraisal capability, as indexed by each measure, are systematically associated with neural activity in a priori defined regulatory neural networks.” Unpacking these statements is important for understanding some of the shortcomings of the present paper’s construction/presentation.**

These comments seem to miss a concept that is central to research on emotion, but is often

unappreciated by those in affective neuroscience - the idea of coherence between different channels of emotional response. As described in reviews by Cacioppo and colleagues, and more recently by Mauss and Gross (and as unpacked in some nice empirical papers from Mauss's lab), emotional coherence refers to the correlation experiential, behavioral and physiological channels of emotional responding. It turns out that ~50% of the time, these channels don't cohere. That is, experiential self-reports, physiological and/ behavioral measures (e.g. facial expressions) only correlate with one another about 50% of the time. There seem to be numerous reasons why this might be the case (e.g. different time scales for different measurements, differing processing systems underlying each measure, different covariates having differential influence over each measure etc.).

Against this backdrop, it could be very interesting to frame the present paper as introducing a different version of the coherence question - do so-called trait measures of reappraisal, self-report-based indices to task-based success and measures of activity in brain systems intercorrelate? Do they cohere? To properly unpack this question, the present paper would need to present a theoretical analysis of why we would or would not expect these measures to correlate/cohere - akin to what has been done in the broader coherence literature. What does measure reflect? To what extent are they thought to reflect the operation of similar vs. different underlying processes/mechanism? Etc. The present paper does not do this.

Response:

We thank the reviewer for this highly productive reframing suggestion, which influenced more changes to the discussion. We were also able to perform some follow-up analyses, which might further inform this point.

Our data call was open to all studies which included a questionnaire measure of reappraisal capability, defined as indicators relating to effective implementation of reappraisal. We consciously permitted questionnaires such as the ERQ, as they are commonly used in the literature as such an indicator, despite strictly referring to reappraisal frequency. The often implicit assumption appears to be that implementation ability and behavioral frequency are strongly functionally coupled and we found it essential to relate to this literature. Therefore, we first briefly added important references supporting these assumptions to the introduction:

"Prior work has linked these three indices with each other, and with prefrontal responses during reappraisal^{12,15–18}"

For concision, we have retained only the most relevant citations in the revised manuscript, but provide the broader set of references motivating this assumption here in the response letter (Drabant et al., 2009; McRae et al., 2012; Kanske et al., 2012; Silvers & Guassi Moreira, 2019; Keller et al., 2021; Kang et al., 2026).

The fact that all available questionnaires focused on beliefs about habitual behavior, rather than beliefs about implementation ability, was very surprising to us, but can also be taken as further evidence that most fMRI researchers believe these two aspects are linked.

In discussing the reviewer's comment with the consortium, it appeared that three large studies (N=444) measured an additional version of the ERQ with an instruction framed towards implementation ability. Fixed-effects analyses on these samples confirmed that beliefs about habits and beliefs about implementation are strongly related. Using the ability ERQ version to repeat the trait questionnaire analyses again supported the null findings.

As our results now include these questionnaires, we first changed the introduction to refer to trait questionnaires more generally (instead of habit trait questionnaires specifically):

“We focused on three widely used indices often assumed to reflect related facets of reappraisal capability: trait questionnaires, task-based affective ratings following reappraisal, and amygdala down-regulation during reappraisal.”

In the results, we dedicated the following paragraph to these analyses:

“All studies employed trait questionnaires which mainly capture beliefs about habitual reappraisal use. While relations to implementation ability, as measured by experiments, are theoretically and empirically expected^{15–17}, trait questionnaires directly measuring beliefs about such abilities might show stronger associations. Three large studies additionally administered modified versions of the Emotion Regulation Questionnaire capturing such beliefs about abilities¹⁹. Fixed effects meta-analysis confirmed that beliefs about habitual use and implementation ability are strongly related ($r=.67$, $p<.001$, 95% CI=[0.62, .72], $k=3$, $N=444$). Supporting the null results for habit questionnaires, there was substantial evidence against an association between these ability trait questionnaires and task-based affective ratings ($r=.05$, $p=.275$, 95% CI=[-.04, .15], $BF_{10}=0.13$, $k=3$, $N=444$). In contrast, the association between trait ability questionnaires and amygdala down-regulation was statistically significant, but in the other direction, meaning beliefs in higher reappraisal capabilities were related to less effective down-regulation of the amygdala ($r=-.12$, $p=.010$, 95% CI=[-.21, -.03], $BF_{10}=4.68$, $k=3$, $N=444$). This represents strong evidence against an effect in the expected direction (one-sided $BF_{10}=0.05$).”

In the discussion, we further included the following paragraph to introduce the coherence problem. We further had the same comment in mind when adding the paragraphs concerning degeneracy and task design (R4-1.3, R4-5).

“The lack of convergence challenges a central assumption in affective neuroscience and clinical neuroimaging: that questionnaires, task-based affective change and neural responses provide complementary windows onto a shared underlying regulatory capability, albeit with unique aspects. This matters because the extent to which common reappraisal task findings inform claims about emotion, psychopathology and intervention targets may be overstated if individual indices capture only distinct aspects of regulatory functioning. Indeed, research on emotional coherence shows that experiential, behavioral and physiological channels often diverge because they operate on different time scales and are shaped by different causal influences and biases²⁷. Trait questionnaires reflect recollected beliefs about reappraisal habits and ability across diverse real-life contexts; task-based affective ratings capture momentary regulatory effects under constrained experimental conditions; and neural responses may additionally reflect processes outside conscious awareness²⁸, while being shaped by neurovascular and scanner-related factors unrelated to regulation. The present findings therefore do not merely show weak measurement convergence; they suggest that common indices of reappraisal capability, in fact, may carve the construct at different joints.”

Drabant, E. M., McRae, K., Manuck, S. B., Hariri, A. R. & Gross, J. J. Individual Differences in Typical Reappraisal Use Predict Amygdala and Prefrontal Responses. *Biological Psychiatry* 65, 367–373 (2009).

Kang, T. et al. High-Intensity Transcranial Alternating Current Stimulation for Emotion Regulation and Attentional Bias in Adolescents With Nonsuicidal Self-Injury: A Randomized Clinical Trial. JAMA Network Open 9, e2617336 (2026).

Kanske, P., Heissler, J., Schönfelder, S. & Wessa, M. Neural correlates of emotion regulation deficits in remitted depression: The influence of regulation strategy, habitual regulation use, and emotional valence. Neurolmage 61, 686–693 (2012).

Keller, M. et al. fMRI Neurofeedback-Enhanced Cognitive Reappraisal Training in Depression: A Double-Blind Comparison of Left and Right vLPFC Regulation. Frontiers in Psychiatry 12 (2021).

McRae, K., Jacobs, S. E., Ray, R. D., John, O. P. & Gross, J. J. Individual differences in reappraisal ability: Links to reappraisal frequency, well-being, and cognitive control. Journal of Research in Personality 46, 2–7 (2012).

Silvers, J. A. & Guassi Moreira, J. F. Capacity and tendency: A neuroscientific framework for the study of emotion regulation. Neuroscience Letters 693, 35–39 (2019).

R4-C8:

— **Wrapped up in this issue seems to be an assumption that reappraisal ‘capability’ is a clearly defined psychological concept/construct. At different points the paper treats scores on self-report questionnaires of reappraisal tendencies as well as reappraisal-related changes in task-based self-reports of emotion and brain activity as indicators of ‘capability.’ For example, in the discussion the authors write, “These findings suggest that the three common indices of reappraisal capability capture qualitatively different facets of reappraisal, despite often being interpreted interchangeably. The absence of a common core among these indicators challenges the assumption that they reflect a shared underlying capacity.”**

Here it would be essential for the authors do define what ‘capability’ means. The ERQ is typically seen as measuring beliefs about *tendencies* to engage in reappraisal. Changes in self-reports of affect might measure capability - in a lab setting, when the experimenters tells people when and how to regulate. [...]

Response:

The reviewer raises an important point, closely linked to the previous comment (R4-C7). The changes made according to this previous comment also aimed to partially address the present comment (R4-C8). This included references concerning the assumed link between these levels of analysis (including habits and experimental implementation), the expansion to the ability ERQ version, and a closer discussion concerning how the three levels of analysis relate to each other.

Moreover, we added our definition of reappraisal capability to the introduction:

“In this context, reappraisal capability refers to individual differences in whether, and how effectively, people use reappraisal to reduce negative affect.”

In the discussion, we tried to further clarify the meaning of the different modalities:

“Trait questionnaires reflect recollected beliefs about reappraisal habits and ability across diverse real-life contexts; task-based affective ratings capture momentary regulatory effects under constrained experimental conditions; and neural responses may additionally reflect processes outside conscious awareness²⁸, while being shaped by neurovascular and scanner-related factors unrelated to regulation.”

Further, the discussion paragraph provided in R4-C5 included changes motivated by the present comment:

“[...] On the one hand, previous work showed associations between reappraisal habits, ability, and neuro-behavioral implementation^{15–17}, including causal evidence that neuromodulation affects reappraisal habits¹⁸. Similarly, we found a very strong correlation between habit and ability questionnaires, pointing towards a functional relation on a conceptual level. On the other hand, standard fMRI tasks bypass much of the naturalistic reappraisal sequence: participants are instructed when and how to regulate, receive practice trials and respond to standardized stimuli optimized for reliable group-level effects. Therefore, individuals with higher reappraisal capability might derive their real-world advantage less from universally superior neural implementation than from selecting, initiating and sustaining reappraisal in personally meaningful contexts, and from encountering environments that more readily afford successful reappraisal³⁶. This fits broader arguments that self-regulation is situated rather than governed by a context-general regulatory resource³⁷. Future paradigms should therefore additionally engage other regulatory stages including strategy selection, personally relevant stimuli and spontaneous deployment.”

Notably, we did not intend to suggest that the three types of capability indices can be reduced to a single variable but rather that covariance is created by functional interactions between these levels. This includes the common assumption of a shared capability, while unique aspects are expected as well for each component (analogous to factor analysis).

R4-C9:

[...] And changes in amygdala response - which once might have been taken to reflect changes in ‘emotional response’, broadly construed, are now not seen by many to be that. Instead, the amygdala is now seen more as a neuromodulatory region that influences the focus of attention, speed and fidelity of perception and consolidation of information into memory. Few affective neuroscientists today think its activity **should correlate with self-reports of emotion. Instead, current approaches look to multivariate approaches to define neural signatures that have been validated to correspond to variations in self-reported emotional states (e.g. the PINES signature from Chang et al).**

With this in mind, finding such as, “...participants with stronger down-regulation of the amygdala via cognitive reappraisal also showed a better down-regulation of task-based affective ratings. Yet, the effect size was comparatively small....” don’t really land with this reader, because they seem to depend on a kind of ‘straw man’ set of naive assumptions about what these different measures reflect.

Response:

Concerning the reviewer's comments on the role of the amygdala, we respectfully agree with several but not entirely with all points being made.

To elaborate, we fully agree with the complex and more domain-general functional description of the amygdala and included these in the manuscript. Still, there are three major arguments, which in our opinion make a test of the amygdala's contribution to emotion regulation capabilities outstandingly important: (1) the amygdala remains one of the most important/frequently studied neural targets in basic and applied affective neuroscience, including neuromodulation for emotional problems, as discussed under R4-2. Especially the recent trend towards focused ultrasound targeting the amygdala in mental disorders, published in major journals, highlight that anatomically localized hubs still play a major role in affective neuroscience on well-being (e.g., Algermissen et al., 2026, Barksdale et al., 2025, Fonzo et al., 2026). Exactly this dissonance between (a) the amygdala's practical relevance in these fields and (b) skepticism stemming from basic affective neuroscience (which we largely share, see e.g. Sicorello & Schmahl, 2021) motivated us to provide a well-powered test for the amygdala's role in reappraisal capabilities to help resolve this tension and inform the larger field based on well-powered evidence. Even within the consortium, both perspectives on the amygdala are highly prevalent, to us highlighting the need for a strong/decisive test of the amygdala as an indicator of reappraisal capability. (2) The relatively large-scale efforts developing affective neural signatures often showed a within-person association between amygdala responses to negative stimuli and self-reported negative affect in the scanner, when both are measured at the same time. While these correlations are smaller than for whole-brain decoding, single regions have lower reliability than whole-brain signatures (Kragel et al., 2020; Sicorello et al., 2021). Therefore, smaller effect sizes should be expected and can still be theoretically meaningful, although predictive utility is severely limited. (3) Crucially, even in the present study, we found a significant preregistered between-person association between amygdala down-regulation and task-based affective reports, further supporting its relevance. The amygdala finding was one among a very restricted number of regions, further supporting its relevance based on our large dataset, while neural signatures like the PINES did not show such associations.

We did not intend to note the small effect size to argue against the amygdala's relevance. Rather, we followed up the effect size statement with a power analysis to show how large the discrepancy is between needed and actual sample sizes in the field: "To detect this effect size with a statistical power of 80%, a single study would need to test 1105 participants". This indicates a lack of statistical power for most individual difference fMRI studies, also indicated by large-scale studies in other domains (Winter et al., 2022; Elliott et al., 2020; Marek et al., 2022; Han et al., 2022; Sicorello et al., 2025). If uncorrected, this leads to an excess of false positives in the literature, which we aim to help prevent with our study. Previously, the classical dlPFC-amygdala model of reappraisal capability appeared to have considerable support, based on several small studies reporting positive findings, highlighting this issue (Morawetz & Basten, 2024).

We agree that whole-brain multivariate patterns such as the PINES achieve higher within-person predictive accuracy than the amygdala alone — likely due to the multi-component nature of mental events and the larger reliability of distributed patterns. However, for theory-driven anatomical inference, model weights from machine learning patterns are not directly interpretable as neural activation (Haufe et al., 2014); source reconstruction is required, which introduces an additional abstraction step and is otherwise statistically similar to our approach (i.e., one involves direct voxel-outcome correlations while the other involves voxel-prediction correlations). Our voxel-wise correlation approach was specifically chosen for theory-guided

anatomical inference about reappraisal capabilities, mirroring large parts of neuroscientific research on well-being.

In sum, the amygdala's role is contested in specialized circles but remains highly influential in basic research, applied settings, and clinical models of emotion regulation. We believe testing these influential theories with well-powered data — and showing that the amygdala's relevance is nuanced — is central to the paper's contribution. We have revised the relevant Discussion section to acknowledge the contemporary neuroscientific debate:

“Only the association between task-based affective ratings and amygdala down-regulation was statistically significant, yet with an effect size ($r = .08$) so small that more than 1,000 participants would be needed for 80% power — far exceeding typical sample sizes in the field.”

“For task-based affective ratings, we found associations in emotion-generative rather than prefrontal emotion-regulatory regions: the amygdala, the caudal pons, and — when excluding a Covid-era outlier dataset — sensorimotor and dorso-parietal attention networks, outside of canonical emotion regulation regions. The amygdala’s role is not uncontroversial: while it tracks concurrent negative affect in well-powered studies^{29,30}, it also responds to positive and otherwise salient stimuli^{7,31}. We further confirmed with four large samples that it is not consistently modulated by reappraisal²⁶. The amygdala is better understood as a neuromodulatory hub facilitating attention, memory encoding and consolidation³², and allostatic responding — processes relevant to emotion generation but not specific to it.”

Algermissen, J., Rascu, M., Weber, L. A., den Boer, T., Martin, E., Treeby, B., ... & Klein-Flügge, M. C. (2026). Low-intensity focused ultrasound to human amygdala reveals a causal role in ambiguous emotion processing and alters local and network activity. *Neuron*, 114(7), 1269-1289.

Barksdale, B. R., Enten, L., DeMarco, A., Kline, R., Doss, M. K., Nemeroff, C. B., & Fonzo, G. A. (2025). Low-intensity transcranial focused ultrasound amygdala neuromodulation: a double-blind sham-controlled target engagement study and unblinded single-arm clinical trial. *Molecular psychiatry*, 30(10), 4497-4511.

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., ... & Hariri, A. R. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological science*, 31(7), 792-806.

Fonzo, G., Enten, L., DeMarco, A., Kline, R., & Nemeroff, C. (2026). Single-Session Focused Ultrasound Modulation of Amygdala Emotional Reactivity and Resting State Connectivity: A Double-Blind, Sham-Controlled Study in Healthy Individuals and Mood, Anxiety, and Trauma-Related Disorders. *Biological Psychiatry*, 99(10), S12-S13.

Han, X., Ashar, Y. K., Kragel, P., Petre, B., Schelkun, V., Atlas, L. Y., ... & Wager, T. D. (2022). Effect sizes and test-retest reliability of the fMRI-based neurologic pain signature. *NeuroImage*, 247, 118844.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87, 96-110.

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2020). fMRI can be highly reliable, but it depends on what you measure. *PsyArXiv*.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., ... & Dosenbach, N. U. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902), 654-660.

Morawetz, C. & Basten, U. (2024). Neural underpinnings of individual differences in emotion regulation: A systematic review. *Neurosci. Biobehav. Rev.* 162, 105727.

Sicorello, M., Herzog, J., Wager, T. D., Ende, G., Müller-Engelmann, M., Herpertz, S. C., ... & Niedtfeld, I. (2021). Affective neural signatures do not distinguish women with emotion dysregulation from healthy controls: A mega-analysis across three task-based fMRI studies. *Neuroimage: Reports*, 1(2), 100019.

Winter, N. R., Leenings, R., Ernsting, J., Sarink, K., Fisch, L., Emden, D., Blanke, J., Goltermann, J., Opel, N., Barkhau, C., Meinert, S., Dohm, K., Reppe, J., Mauritz, M., Gruber, M., Leehr, E. J., Grotegerd, D., Redlich, R., Jansen, A., ... Hahn, T. (2022). Quantifying Deviations of Brain Structure and Function in Major Depressive Disorder Across Neuroimaging Modalities. *JAMA Psychiatry*, 79(9), 879–888. <https://doi.org/10.1001/jamapsychiatry.2022.1780>

R4-C10:

— Along these lines, the authors might want to think about whether there is something new and interesting their data can say about so-called trait tendencies vs. task-based measures. For example, the authors report that, “There was no significant association between trait questionnaires and either outcome, i.e., amygdala down-regulation or task-based affective ratings.” And that, “There was no statistically significant association between trait questionnaires and neural activity during reappraisal.”

On one hand, it would be great to have these findings published in the literature. They address a file-drawer problem for work on emotion regulation where many who have done reappraisal studies - and have included the ERQ - have known for years that ERQ scores do not predict either changes in task-based changes in self-reports of affect or changes in amygdala or other brain correlates of reappraisal. To date, however, no one has published these failures to see such correlations.

On the other hand, once noted, the key question is WHY ‘trait’ measures, like the ERQ, predict good outcomes in everyday life but do NOT correlate with task based measures of reappraisal-related changes in emotion. The present analyses and discussion do not provide an answer. One way to look for answer could involve drawing on variants of ‘extended process models’ of emotion regulation (see, e.g. examples from Gross, Sheppes, Ochsner and colleagues) positing that the implementation of a strategy like reappraisal comes only after a series of other processing stages have occurred. These stages include, critically, an initial decision about whether reappraisal (or some other strategy) is called for at the present moment, and then if regulation is to happen, a decision about what strategy to use. It is very likely that ‘trait’ measures like the ERQ pick up individual variability in these decision/selection processes as much, if not more, than they pick up on variability in the fidelity with which individuals can implement specific reappraisals when told when and how to do so. The latter stage - implementation - all that current fMRI studies tap into.

Response:

We thank the reviewer for identifying this as the central explanatory gap in the manuscript as well as pointing out the relevance of the file drawer problem. As noted in the previous comment, the literature still strongly suggests associations between these levels across cross-sectional and neuromodulation studies. We believe that our manuscript adds to the literature by empirically evaluating predictions from what are arguably misguided models and popular beliefs about the neurobiology of emotion regulation. Notably, the association with the amygdala did conform to predictions, while also confirming previous large-scale studies suggesting that between-person associations with neuroimaging are mostly very small (Marek et al., 2022). We propose a two-level account of why trait and task-based measures are expected to dissociate, the first focusing on the extended process model stages and the second on stimuli optimized for naturalistic/personalized relevance.

This is now addressed in two paragraphs in the Discussion. The Inzlicht et al. literature is explicitly integrated:

“The absence of associations between trait questionnaires and task-based or neural indices points to an explanatory gap that can be understood through process models of emotion regulation^{1,26}. Most trait questionnaires capture habitual tendencies across multiple temporal stages, including encountering and recognizing reappraisable situations, deciding whether to regulate, and selecting reappraisal over competing strategies. On the one hand, previous work showed associations between reappraisal habits, ability, and neuro-behavioral implementation^{15–17}, including causal evidence that neuromodulation affects reappraisal habits¹⁸. Similarly, we found a very strong correlation between habit and ability questionnaires, pointing towards a functional relation on a conceptual level. On the other hand, standard fMRI tasks bypass much of the naturalistic reappraisal sequence: participants are instructed when and how to regulate, receive practice trials and respond to standardized stimuli optimized for reliable group-level effects. Therefore, individuals with higher reappraisal capability might derive their real-world advantage less from universally superior neural implementation than from selecting, initiating and sustaining reappraisal in personally meaningful contexts, and from encountering environments that more readily afford successful reappraisal³⁶. This fits broader arguments that self-regulation is situated rather than governed by a context-general regulatory resource³⁷. Future paradigms should therefore additionally engage other regulatory stages including strategy selection, personally relevant stimuli and spontaneous deployment.”

Notably, the findings concerning the sensorimotor areas fits recently proposed emotion regulation theories and large-scale findings on stress-vulnerability, which we try to highlight more clearly in the revised discussion:

“Our finding of sensorimotor associations is consistent with extended process models that assign a role to sensorimotor cortices and brainstem in the action phase of regulation²⁶. Recent large scale evidence indicates that even in the absence of overt motor responses, sensorimotor cortices carry rich information about emotion generation²⁶, are modifiable by reappraisal strategies²⁶, are related to individual differences in stress vulnerability^{11,33}, and a promising target for neuromodulation in depression³⁴.”

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., ... & Dosenbach, N. U. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902), 654-660.

R4-C11:

— How do the authors make sense of their failure to see any PFC regions associated with individual differences in light of evidence collected in fTMS and neurofeedback studies that individual differences in PFC engagement during reappraisal are important? For example, recent studies of reappraisal in clinical groups seems to show that ventrolateral PFC activity during reappraisal is weaker (De la Pena-Arteaga et al., 2021) and may inversely correlate with symptom severity in MDD and BPD (Keller et al., 2022). Notably, however, in patients with MDD, neurofeedback focused on ventrolateral PFC can increase the effective daily use of reappraisal (Keller et al., 2021). See also recent paper by He et al in J. Neuroscience (2023).

Response:

We see two distinct points in this comment, which we address in turn.

On clinical case-control studies showing PFC deficits: we did not focus on individual clinical studies, because our previous meta-analysis of the published fMRI reappraisal literature in mental disorders shows that 'neural deficits in reappraisal' do not converge on canonical ER networks either within or between disorders (Morawetz et al., 2025). In a prior review specifically on Borderline Personality Disorder, we found that although all studies reported 'prefrontal deficits', the regions identified showed no overlap with each other or with canonical networks (Sicorello & Schmahl, 2021). Individual case-control studies in the literature are additionally often subject to well-documented issues: complex confounding, fuzzy diagnostic boundaries, file drawer bias, low statistical power, and unconstrained researcher degrees of freedom. Large-scale preregistered consolidations of this literature using dimensional more targeted measures — the goal of NIDER — are therefore in our opinion essential for drawing strong anatomical conclusions.

On neuromodulation evidence: the disconnect between our findings and neuromodulation outcomes may reflect the distinction between process-based and between-person research. The Keller et al. (2021) vLPFC neurofeedback study used ERQ scores as a primary outcome — and found improvements in both clinical and healthy control groups, suggesting vLPFC was not a disorder-specific target. This is consistent with our argument that process-based research (which correctly identifies neural systems involved in reappraisal implementation) might provide valid neuromodulation targets, even when between-person studies fail to confirm anatomically specific individual-differences associations. Our study specifically cautions against inferring neuromodulation targets from the between-subject literature, not from process-oriented research. More broadly, the neuromodulation field remains highly heterogeneous in designs, protocols, and anatomical targets (Khaleghi et al., 2025). There is evidence that stimulation is likely effective in depression across a range of targets (Downar & Daskalakis, 2013; Hu et al., 2024). This might suggest that common protocols may work through distributed global effects rather than region-specific mechanisms — further arguing for caution in inferring specific targets from individual-differences data.

We have added this paragraph to the revised Discussion clarifying this distinction:

“These findings might have direct implications for how neural correlates are used to guide clinical inference and neuromodulatory interventions. The absence of prefrontal individual-difference associations does not invalidate process-oriented findings: the neural systems identified in within-person reappraisal research could remain viable targets for neuromodulation, as they reflect the implementation of reappraisal regardless of between-

person variation. However, our results support caution against inferring individual regulatory capacity from regional brain activation — whether in the prefrontal cortex, amygdala, or elsewhere — as is common in clinical neuroimaging models of emotion dysregulation^{6,7}. The success of neuromodulatory interventions targeting these regions^{5,18} may reflect distributed network reorganization rather than correction of a specific regional deficit^{34,43}, consistent with evidence that different stimulation sites can achieve similar clinical outcomes. This fits a systems neuroscience perspective where single regions cannot be interpreted in isolation. In support, recent well-powered studies linked individual differences in affective functioning with reappraisal-contingent connectivity between the amygdala and prefrontal areas (vlPFC, dlPFC, dmPFC)⁴⁴ as well as shifts along primary gradients of macroscale neurofunctional organization⁴⁵. Further testing these approaches represents a natural next step for the NIDER dataset and individual-difference neuroimaging of emotion regulation more broadly.”

Downar, J., & Daskalakis, Z. J. (2013). New targets for rTMS in depression: A review of convergent evidence. *Brain Stimulation*, 6(3), 231–240. <https://doi.org/10.1016/j.brs.2012.08.006>

Hu, Y.-T., Hu, X.-W., Han, J.-F., Zhang, J.-F., Wang, Y.-Y., Wolff, A., Tremblay, S., Hirjak, D., Tan, Z.-L., & Northoff, G. (2024). Motor cortex repetitive transcranial magnetic stimulation in major depressive disorder—A preliminary randomized controlled clinical trial. *Journal of Affective Disorders*, 344, 169–175. <https://doi.org/10.1016/j.jad.2023.10.058>

Khaleghi, A., Samiei, H., Zarafshan, H., Baloochi, S. A., & Mohammadi, M. R. (2025). Effectiveness of fMRI-based Neurofeedback Therapy on Depression: A Systematic Review. *Clinical Psychopharmacology and Neuroscience*, 23(3), 337.

Morawetz, C., Hemetsberger, F. J., Laird, A. R., & Kohn, N. (2025). Emotion regulation: From neural circuits to a transdiagnostic perspective. *Neuroscience & Biobehavioral Reviews*, 168, 105960. <https://doi.org/10.1016/j.neubiorev.2024.105960>

(Point C).

R4-C12:

The paper mentions that currently used scanner-based appraisal tasks may not map onto what people do in everyday life when regulating emotion. That is, current tasks may not be ecologically valid. This is a very interesting idea. A likely very important one. Can the authors say more about what they mean and how this might bear on the findings they observed?

— The above discussion about extended process models might provide some insights on this issue. For example, if current fMRI tasks only measure the implementation of a strategy in a highly constrained context where individuals are told when and how to re-appraise, and are given practice in doing so, then such procedures may get rid of many sources of naturally occurring individual variability that are very meaningful to the full range of individual differences in emotional well-being the authors are interested in. These include factors like those mentioned earlier, including the factors that determine whether an individual chooses to regulate, and if so, how.

Response:

We fully agree with the reviewer's analysis and propose a fuller model in the Discussion. The respective changes are shown under the closely related R4-10.

In brief, they rely on two perspectives; first, the selectivity of common fMRI for specific components of the extended process model; second, the concept of “affordances” as a guiding principle towards individually meaningful tasks.

Notably, R1-C2 stimulated follow-up analyses showing that even in samples with considerable individual differences in affective functioning, most variance in ratings lies on the within-person level, confirming previous arguments that common fMRI tasks on affective functioning are neither designed, nor suitable, for research on well-being. This is despite them being used on thousands of individuals towards this purpose, with findings having had a strong practical influence on clinical science, practice, and public perception of these topics.

R4-C13:

— Another consideration here - and one that bears on the question of coherence across trait versus task- based measures of reappraisal, comes from recent work by Michael Inzlicht and colleagues. Inzlicht has noted that “trait” measures of self-control - that is, self-report questionnaires measuring some aspect of self-control - and “state” measures of self-control - that is, task-based measures of the ability to exert a specific kind of self-control - are often related to different kinds of outcomes. Notably, he argues that the trait-based questionnaires are what predict most of the good things we typically think of when we think about self-control, whereas the task-based measures rarely predict anything terribly meaningful. See his recent review/opinion papers on this. Could this illuminate a discussion of the short-comings of current task-based approaches for studying reappraisal?

Response:

We thank the reviewer for suggesting this literature, which directly parallels the questionnaire-task disconnect in our data. The more comprehensive changes to the manuscript were presented under R4-10 (also touching on R4-1, R4-5, and R4-7) where the following middle sentence specifically regarded two crucial studies from Michael Inzlicht's lab:

“Therefore, individuals with higher reappraisal capability might derive their real-world advantage less from universally superior neural implementation than from selecting, initiating and sustaining reappraisal in personally meaningful contexts, and from encountering environments that more readily afford successful reappraisal³⁶. This fits broader arguments that self-regulation is situated rather than governed by a context-general regulatory resource³⁷. Future paradigms should therefore additionally engage other regulatory stages including strategy selection, personally relevant stimuli and spontaneous deployment.”

Concerning the relation to relevant external outcomes, we also want to highlight part of our response to R1-C1

“The reviewers' suggestion that future work should characterize what each measure uniquely contributes is one we strongly endorse. We believe this will require external validation criteria that do not trivially overlap in modality with any of the three measures (e.g., another questionnaire), with clinical outcomes in diagnostically characterized transdiagnostic samples representing one especially promising direction. At the same time, we believe our second goal — relating these three indices to canonical regulation networks — is also informative in this

context, as none of the three indices showed the hypothesized associations that are often assumed in the literature. As the reviewers suggested, we believe a natural next candidate is clinical outcomes in transdiagnostic samples, which represents the next long-term phase of the NIDER consortium.”

R4-C14:

— is there any potential issue associated with the fact that nearly of the 41 reappraisal studies included in the current analysis used photographs as their stimuli? Assuming most of these come from the international effective picture system, or some similar set, then these images were created to as to elicit affective responses that are normatively the same across individuals. Could the use of such stimuli actually work against finding meaningful individual differences?

Response:

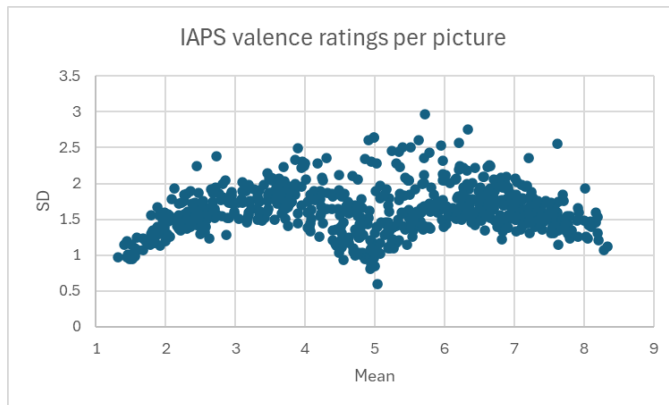
The reviewer raises a critically important issue. IAPS-like pictures have dominated the task-based affective fMRI literature, especially for research on individual differences, mirrored in their representation within our study. We agree with the reviewer’s criticism, especially on the grounds of matching naturalistic emotion processes with relevance to participants:

“Future paradigms should therefore additionally engage other regulatory stages including strategy selection, personally relevant stimuli and spontaneous deployment.”

We would like to add a nuance within our response. As IAPS pictures were designed to elicit strong group responses, they still can exhibit substantial individual differences in ratings. Most picture-wise standard deviations ranged between 1 and 2, based on the original IAPS norms (see figure below). The average SD was 1.6, indicating that, on average for all pictures, approximately 95% of participants were spread across 6.4 points of the original 1-9 Likert scale, which represents considerable individual differences. Empirically, there are strong gender effects for IAPS ratings as well as associations with personality traits (Aluja et al., 2015). Decoding individual differences in IAPS-related ratings from fMRI responses leads to moderate-to-large effect sizes (Sicorello et al., 2025) based on empirical effect size norms for individual differences (Gignac & Szodorai, 2016). Together, this indicates that responses to IAPS pictures exhibit substantial, meaningfully explainable individual differences.

Notably, simply changing the stimulus modality from pictures to videos does not appear to resolve this issue: our dataset includes studies using video stimuli, and we observe no meaningful difference in between-person effect sizes compared to photograph-based studies, with between-study heterogeneity remaining low across all analyses. This might suggest the problem lies in the personal relevance dimension, not in the specific modality or format of stimuli. As noted within our addition to the Discussion, future paradigms using ideographically selected stimuli — whether personally relevant images, film clips, or language-based triggers — would be, in our opinion, substantially better positioned to detect the between-person associations that currently evade detection with standardized stimuli.

Additionally, we believe that, theoretically, even stimuli which uniformly increase negative affect across individuals might be viable, if people show individual differences in how they regulate back down from such an affective peak.



Aluja, A., Rossier, J., Blanch, Á., Blanco, E., Martí-Guiu, M., & Balada, F. (2015). Personality effects and sex differences on the International Affective Picture System (IAPS): A Spanish and Swiss study. *Personality and Individual Differences*, 77, 143-148.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>

Sicorello, M., Gianaros, P. J., Wright, A. G. C., Petre, B., Kraynak, T. E., Manuck, S. B., ... & Wager, T. D. (2026). The functional neurobiology of dispositions towards negative emotions. *Nature Communications*, 17(1), 5622.

Reviewer comment 4.1:

“The first [comment] concerns the characterization of the ERQ. I know this measure well. It does not measure 'capability'. It measures self-reported tendencies to engage in specific patterns of thinking that characterize reappraisal. I do not believe it is appropriate to characterize it as a capability measure. The authors have revised the paper to provide a definition of capability that is broad enough to nod in this direction. I suppose I wish the authors did a bit more to explicitly acknowledge that many in the field see this as a tendency measure”

Response:

We were very grateful for the kind words of reviewer 4 concerning the revised version. We fully agree with the point raised here, and we thank the reviewer for pressing us on this issue. We share the view that the ERQ (and comparable instruments) index self-reported *tendencies* to reappraise rather than the *ability* to implement reappraisal. Our intention was never to equate the two: the broad definition of "capability" was meant to span tendency, ability and neural implementation as distinct but related indices, not to recast the ERQ as an ability measure. To remove any ambiguity we now state this explicitly and early in the Introduction, while retaining, in the Results ("Divergence of capability indices") and Discussion ("The trait–task gap"), our fuller rationale for including tendency questionnaires: their widespread use and their theoretically and empirically proposed functional relation to implementation ability, consistent with the strong habit–ability correlation we report. Notably, the word “capability” was exchanged with “capacity” throughout the manuscript to provide consistency with the new title suggested by the editorial board. The sentence added to the introduction is:

“Notably, the most widely used trait questionnaires are generally understood to index self-reported habitual tendencies to reappraise rather than the ability to implement reappraisal. We nonetheless included them given their proposed functional link to such ability^{15–17} and their widespread use”

Reviewer 4.2:

“The second concerns how we might interpret some of what authors report - like ERQ scores positively predicting negative affect-related DVs. We tend to think that when we try regulating that all DVs reflecting affective responding should go down. This paper suggests that isn't always the case and that different DVs don't really track with each other. One wonders is two things could be doing on here. One is that reappraisal could lead people not to simply diminish emotion but to change it. To the extent that any DV reflects arousal/intensity of emotion rather than the quality/valence of emotion (e.g. the amygdala!) then we might not always expect its activity to go *down* when we reappraise. A second question is whether some people tend to try reappraisal, in everyday life or the lab, because they tend to have strong (negative) emotions. Their emotions might be so strong that - despite their efforts - they try and fail. The could be because the emotions were too strong. It could also be because reappraisal might not have been the best strategy to use then (See papers on reappraisal choice, e.g., or EMA Studies, suggesting we tend not to choose reappraisal when emotions are strong), or that they aren't very good at it (for whatever reason). There are examples in the literature (e.g. papers from Mauss's lab about reappraisal not always being the best thing to do), including clinical examples (e.g. recent paper from Herzog et al with depressed individuals) consistent with this idea”

Response:

We thank the reviewer for these generative interpretations, both of which we now incorporate. (i) On quality versus intensity, we agree that reappraisal may alter the quality/valence of a response without proportionally reducing its salience, consistent with our finding that the amygdala — an arousal/salience-related neuromodulatory hub — is not consistently down-modulated by reappraisal. (ii) On selection, choice and reactivity, we now note that a positive tendency–negative-affect association need not imply failure: habitual reappraisers may be individuals with stronger reactivity, and intense emotions are both less amenable to and less likely to be selected for reappraisal, so attempts may be initiated when least likely to succeed. We connect this to the reappraisal-choice and person-by-situation literatures and to clinical evidence that spontaneous reappraisal engagement does not straightforwardly track reduced negative affect in depression (Herzog et al., 2025), which we believe strengthens our situational-affordances argument. The additions in response to the two points are shown below (respectively), marked in italics:

“We further confirmed with four large samples that it [the amygdala] is not consistently modulated by reappraisal²⁶. The amygdala is better understood as a neuromodulatory hub facilitating attention, memory encoding and consolidation³², and allostatic responding – processes relevant to emotion generation but not specific to it. *One possibility is that reappraisal changes the quality or valence of an emotional response without proportionally reducing its salience.*”

“Therefore, individuals with higher reappraisal capacity might derive their real-world advantage less from universally superior neural implementation than from selecting, initiating and sustaining reappraisal in personally meaningful contexts, and from encountering environments that more readily afford successful reappraisal³⁶. *Relatedly, higher self-reported reappraisal tendency may partly reflect greater regulatory need rather than greater regulatory success: individuals experiencing stronger negative emotions may have more occasion to engage in reappraisal³⁷. At the same time, highly intense negative emotions may render reappraisal less suitable than alternative regulation strategies³⁸, relating reappraisal success to situational demands and individual differences in emotional reactivity^{4,12}.* This fits broader arguments that successful self-regulation *reflects complex person-situation interactions* rather than being governed by a context-general regulatory resource³⁸”