Stage 2 Registered Report

Psychological Science
2025, Vol. 36(5) 367–383

DOI: 10.1177/09567976251335567
www.psychologicalscience.org/PS

Preregistered Direct Replication and Extension of "The Wisdom to Know the Difference: Strategy-Situation Fit in Emotion Regulation in Daily Life Is Associated With Well-Being"

Emily C. Willroth[1], Gerald Young[2], Brett Q. Ford[3], Allison Troy[4], Dorota Swierzewicz[5], and Iris B. Mauss[5]

[1]Department of Psychological and Brain Sciences, Washington University in St. Louis; [2]Psychology Department, University of Ohio; [3]Department of Psychology, University of Toronto; [4]Popular Comms Institute, Lancaster, Pennsylvania; and [5]Department of Psychology, University of California, Berkeley

Abstract

Certain emotion-regulation strategies (e.g., reappraisal) are associated with better well-being and are therefore seen as adaptive (health-promoting) strategies. However, it is unlikely that any strategy is adaptive regardless of context. Indeed, reappraisal is associated with positive outcomes in the context of uncontrollable life stress but negative outcomes in the context of controllable life stress. It follows that individuals who have better "strategy-situation fit" (use reappraisal more during uncontrollable vs. controllable situations) should have better well-being beyond their habitual reappraisal use. A previous test of this hypothesis found that strategy-situation fit in daily life was associated with greater well-being ($N = 74$). We conducted a well-powered preregistered direct replication of this study in 285 U.S. adults. We failed to replicate the original findings and found no evidence for the strategy-situation fit hypothesis, including when accounting for key confounders and moderators. We discuss implications for theory and future research.

Keywords

context, emotion regulation, reappraisal, replication, strategy-situation fit

Received 2/16/18; Revision accepted 3/21/25

Emotion-regulation strategies are sometimes likened to tools in a socioemotional toolbox—people have a variety of tools to choose from, and some tools are better than others for promoting well-being. Following this view, past research on emotion regulation has largely focused on identifying adaptive (i.e., health-promoting) and maladaptive (i.e., health-harming) strategies. For example, reappraisal (i.e., reframing one's thoughts about a situation to change its emotional impact; Gross, 2002) is often seen as a particularly useful tool because it is generally associated with positive outcomes (Gross & John, 2003; Troy et al., 2010). But just as a hammer is not very useful for turning a screw, and a screwdriver has little utility for hammering a nail, the usefulness of emotion-regulation tools may depend on the context in which they are used.

The role of context in determining the adaptiveness of particular emotion-regulation strategies is a topic of great theoretical and practical importance yet has been examined only relatively recently compared with emotion-regulation research more generally (Aldao, 2013; Troy et al., 2013). The controllability of the situation is a particularly important contextual feature to consider

Corresponding Author:
Emily C. Willroth, Department of Psychological and Brain Sciences, Washington University in St. Louis
Email: emily.w@wustl.edu

because it fundamentally determines the utility of efforts to regulate one's own responses to the situation (Folkman, 1984). When a stressful situation is relatively controllable (when the person can change the situation), it may be better to take action to improve the situation itself. When a stressful situation is relatively uncontrollable (when the person can primarily change their reaction to the situation but cannot change the situation itself), it may be adaptive to focus on changing what can be controlled—one's emotions (Collins et al., 1983; Lazarus, 1993). Previous research has found that the outcomes associated with reappraisal depend on the controllability of life stress (Troy et al., 2013), consistent with this idea. Specifically, the ability to use reappraisal was associated with fewer depressive symptoms for individuals with relatively uncontrollable life stress but more depressive symptoms for individuals with relatively controllable life stress.

Given that reappraisal appears to be most adaptive when used in relatively uncontrollable contexts and less adaptive or even harmful when used in relatively controllable contexts, individuals who have good strategy-situation fit (i.e., use reappraisal more during uncontrollable vs. controllable situations) should have better well-being above and beyond their habitual reappraisal use. Indeed, a meta-analysis found that strategy-situation fit was associated with better psychological adjustment (Cheng et al., 2014). However, previous measures of strategy-situation fit have assessed broad, heterogeneous categories of coping (i.e., emotion- vs. problem-focused) rather than specific theoretically grounded emotion-regulation strategies. These studies have also largely relied on one-time self-report measures that ask participants to respond to hypothetical situations. These existing measures of strategy-situation fit are subject to limitations (e.g., limits to validity; Kashdan & Rottenberg, 2010).

The first study that tested the strategy-situation fit hypothesis with a specific theoretically grounded emotion-regulation strategy (i.e., cognitive reappraisal; Haines et al., 2016) provided a strong test of the hypothesis by measuring strategy-situation fit as the within-persons association between reappraisal use and the controllability of situations in daily life using ecological momentary assessment (EMA). Haines et al. found that individuals who used reappraisal more in situations that were perceived as relatively lower in controllability and less in situations that were perceived as relatively higher in controllability had greater well-being. Although this finding offers the strongest support for the strategy-situation fit hypothesis to date and contributes to the understanding of the role of context in the adaptiveness of particular emotion-regulation strategies, the study sample was too small (N = 74) to examine potential key confounders (e.g., habitual reappraisal use) and moderators (e.g., gender). Therefore, replication is needed to build on this original finding while also testing its robustness and generalizability. Results from these tests will have important implications for understanding what constitutes adaptive emotion regulation, how emotion regulation contributes to well-being, and how best to inform clinical interventions and prevention programs.

For these reasons, we conducted a preregistered replication of Haines et al. (2016) in which the original methods were replicated as directly as possible in a well-powered sample more than three times the size of the original sample (original N = 74; replication N = 285). We also extended the original study in four key ways. First, strategy-situation fit is assumed to be a temporally stable individual difference. To test this assumption, we collected an additional week of EMA surveys in a subsample of participants to calculate test–retest reliability of strategy-situation fit across 2 weeks.

Second, we accounted for four potential key confounders. According to the strategy-situation fit hypothesis, strategy-situation fit should be associated with greater well-being above and beyond average strategy use. To test this, we statistically controlled for habitual reappraisal use (people's self-reported tendency to use reappraisal to change their emotions; Gross & John, 2003). In addition, to test whether the association between strategy-situation fit and well-being is driven by variability in strategy use or variability in situational controllability, we statistically controlled for variability in reappraisal use as well as in controllability. Last, to test whether the association between strategy-situation fit and well-being is driven by desirability response bias, we statistically controlled for individual differences in impression management. We conducted extension analyses controlling for potential key confounders regardless of the results of the direct replication analyses because confounders can lead to spurious associations as well as spurious lack of associations.

Third, we assessed the association between strategy-situation fit and well-being in a subset of the data in which participants reported experiencing a negative or unpleasant event. These analyses accounted for the possibility that low levels of reappraisal use may be particularly prevalent in situations that are not negative or unpleasant. This may result in a potential confound that could weaken the observed associations between strategy-situation fit and well-being. We compared the results using the full sample with the results using the subset of data in which participants reported experiencing a negative or unpleasant event.

Fourth, we tested the generalizability of the association or lack of association between strategy-situation fit and well-being across gender (men compared with women) and student status (undergraduate participants compared with community participants). The small sample size in the original study did not allow for moderation tests. The adaptiveness of strategy-situation fit should be universal: Reappraisal use should be most associated with positive outcomes in uncontrollable contexts and least associated with positive outcomes in controllable contexts for everyone. However, it is possible that the association between strategy-situation fit and well-being depends on individual values that vary across gender and student status. Therefore, tests of generalizability are needed.

We hypothesized on the basis of the original finding that (a) better strategy-situation fit would be associated with greater well-being; (b) strategy-situation fit would be relatively stable across 2 weeks; (c) the association between strategy-situation fit and well-being would hold above and beyond habitual reappraisal use, variability in reappraisal use, variability in controllability, and impression management; (d) the association between strategy-situation fit and well-being would be equivalent or stronger when examining only observations in which a negative or unpleasant event was experienced; and (e) the direction of the association between strategy-situation fit and well-being would be consistent across groups (men compared with women, undergraduate participants compared with community participants).

Research Transparency Statement

General disclosures

Conflicts of interest: All authors declare no conflicts of interest. **Funding:** No external funding was used for this project. Gift cards for community participants were purchased with departmental funds. **Artificial intelligence:** At the recommendation of the STAR editor team, we used ChatGPT as a resource to make some of our code more parsimonious. In all instances, we double-checked the accuracy of our code by comparing the approach ChatGPT recommended to our original approach. No other AI-assisted technologies were used in this research or the creation of this article. **Ethics:** This project was approved by the institutional review board at the University of California, Berkeley (Protocol No. 2016-02-8400).

Study disclosures

Preregistration: This study is a registered report. The research aims, hypotheses, methods, and analysis plan were preregistered prior to data collection. The Stage 1 registered report is publicly available (https://osf.io/ap2ws). There were major and minor deviations from the preregistration (for details, see Table 1; Willroth & Atherton, 2024). **Materials:** All study materials are publicly available (https://osf.io/nk5xy). **Data:** De-identified data sets are publicly available (https://osf.io/u3sbr). These include observations for all participants and variables used in the analyses reported herein. **Analysis scripts:** All analysis scripts are publicly available (https://osf.io/gycfv). **Computational reproducibility:** The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

Method

We communicated with the authors of the original study about our plan to conduct a preregistered replication study. They shared the original study materials and confirmed that our procedures were consistent with those of the original study. This project was approved by the institutional review board at the University of California, Berkeley (Protocol No. 2016-02-8400).

Participants

Study advertisements invited individuals to participate who were either "comfortable" or "fearful" of social situations, consistent with the original study and to maximize variability in well-being. However, in line with the original study, this factor was not considered in the analyses. Undergraduate participants received partial course credit for their participation, and community participants received monetary compensation for their participation in the form of gift cards.

The sample size in the original study was 74 (mean age = 23.26 years; 61% women, 58% students). To determine the sample size needed for the replication study, we conducted simulations with the simr package (Green & MacLeod, 2016), an R package designed to estimate statistical power for multilevel models. We used data from the original study to simulate additional data for the power simulations. On the basis of the results of 100 simulations, we aimed to recruit 285 total participants after exclusions. A sample size of 285 has at least 95% power to detect two-way crossover interactions between controllability and well-being (i.e., strategy-situation fit) for all six types of well-being observed in the original study. We defined crossover interactions as interactions in which the effect size observed in the original study was present in one group and present in the opposite direction in the other group.

In addition, a sample size of 285 provides sufficient statistical power to test whether gender or student status moderates effects, that is, to detect three-way

Table 1. Preregistration Deviations

No.	Details		Stage 1 registered report wording	Deviation description	Reader impact
1	Type Reason Timing	Study design Other (please explain) Before data collection	"Participants will complete EMA surveys in Qualtrics. Links will be texted to their smartphones using the experience sampling software Survey Signal."	"Participants completed EMA surveys in SEMA 2 or SEMA 3 (O'Brien et al., 2024)." We initially planned to use Qualtrics rather than SEMA to collect EMA data because SEMA had not yet gone through our institution's data security review. However, after SEMA was approved, and before data collection began, we switched to SEMA for consistency with the original Haines et al. (2016) study.	This deviation increased consistency between the original Haines et al. (2016) study and the replication study.
2	Type Reason Timing	Study design Other (please explain) During data collection	"Participants will then complete an initial laboratory session, in which they will complete demographic questionnaires and well-being measures. They will then be instructed on the EMA procedure, which will take place over one ($N = 165$) or two weeks ($N = 115$)."	Because of the COVID-19 pandemic, initial sessions conducted after March 2020 took place over Zoom. As a result of slower-than-expected enrollment, we also reduced the target sample size for the subsample of participants who completed 2 weeks of the EMA protocol to 84 participants. Both deviations were peer-reviewed and approved via a revision to the Stage 1 manuscript. After exclusions, our final analytic subsample of participants who completed 2 weeks of the EMA protocol was 80 participants.	Because the initial session involved completing only surveys and receiving instructions, it is unlikely to have impacted study results. The smaller than planned sample size for the test–retest reliability analyses may have reduced the precision of the test–retest reliability estimate.
3	Type Reason Timing	Study design Other (please explain) During data collection/ after data access	"EMA surveys every 72 minutes (± 30 minutes) between 10 a.m. and 10 p.m. for 7 days"; "an additional week of EMA surveys in a subsample of participants (target $N = 115$) to calculate test-retest reliability of strategy-situation fit across two weeks"	Unexpected collection errors were identified after data collection was complete. Thirteen participants had duplicate ITERATION values (for more details, see Procedures section). EMA surveys with duplicated ITERATION values were excluded prior to analysis. Forty-two participants completed more than the expected number of EMA surveys. Surveys after the final planned survey were excluded prior to analysis.	Removing the EMA surveys with duplicate ITERATION values is unlikely to have impacted study results because it affected a small number of surveys and participants relative to the sample size. Removing EMA surveys completed after the last planned survey cannot impact study results because participants' experiences after completing the last planned survey cannot retroactively impact their experiences in the planned surveys that occurred earlier in time.

(continued)

Table 1. *(continued)*

No.	Details		Stage 1 registered report wording	Deviation description	Reader impact
4	Type Reason Timing	Study design Other (please explain) During data collection/ after data access	"At the end of the original survey, after all the items from the original study were asked, participants were also asked the binary item 'Have you experienced a negative/ unpleasant event since the last survey?' (No = 0; Yes = 1)."	Fourteen observations of the negative event variable from four participants were collected on a scale from 0 to 100 instead of the intended binary scale and thus were recoded as missing.	Because this affected less than 0.10% of observations, it is unlikely to have impacted study findings.
5	Type Reason Timing	Study design Other (please explain) During data collection/ after data access	N/A	A resolved SEMA bug caused responses to slider questions that were made all the way to the left or all the way to the right to be recoded outside of the range from 0 to 100. We recoded these values to 0 and 100, respectively, in line with recommendations from the SEMA team.	The affected responses were made at the left-most and right-most parts of the visual scale. Thus, the recoded values match the response that SEMA should have recorded originally.
6	Type Reason Timing	Study design Other (please explain) During data collection/ after data access	N/A	For 1.7% of occasions, SEMA recorded the incorrect date value (i.e., 12/31/1969 for a study conducted between 2019 and 2023). Thus, we replaced all recorded dates prior to the start of data collection with missing values, and we were not able to compute lagged reappraisal for these occasions.	Because this affected only 1.7% of observations, it is unlikely to have impacted study findings.
7	Type Reason Timing	Analysis New knowledge After data access	"One potential problem that we anticipate is failure for one or more of the models to converge. If this happens, we will attempt to achieve convergence by changing the optimization parameters (e.g., increasing the maximum number of iterations). If the model still does not converge, we will begin removing predictors in the following order: lagged reappraisal, time."	We rescaled the EMA items by dividing by 100 to achieve model convergence.	Results can be interpreted on a scale from 0 to 1 rather than 0 to 100, but this does not affect results.

Note: EMA = ecological momentary assessment; N/A = not applicable; SEMA = Smartphone Ecological Momentary Assessment.

interactions between controllability, well-being, and gender (men vs. women) and between controllability, well-being, and student status (student vs. nonstudent). Specifically, we had at least 95% power to detect crossover interactions (i.e., an effect in the opposite direction for men compared with women and for students compared with nonstudents). Crossover interactions are of particular theoretical interest because a crossover interaction would provide evidence that strategy-situation fit is harmful (rather than beneficial or inert) for a particular subgroup. This information is especially important for informing clinical intervention and prevention programs. We did not have sufficient power to test for four-way interactions (e.g., between controllability, well-being, gender, and student status) or attenuated moderation (differences in magnitude of effects in the same direction between groups). These types of interactions are not of significant theoretical interest and, based on our power analyses, would require prohibitively large sample sizes (i.e., > 500).

All participants who reported their age (three participants did not report their age) were between 18 and 32 years old, consistent with the original study. We aimed to recruit a replication sample that was 61% women (± 2%) and 58% students (± 2 %), in line with the original sample. We continued to recruit participants until we reached the stated number of participants per subgroup after exclusions. The final analytic sample postexclusions included 13,273 observations from 285 participants (59% women, 38% men, 3% other gender identity; 59% undergraduate students, 41% community participants; mean age = 23.07 years, range = 18–32 years, *SD* = 4.28).

Procedure

Participants first completed a brief online screening questionnaire before they were invited to participate in the study. This allowed us to achieve our sampling goals in terms of proportions of men and women participants and student and community participants. Participants then completed an initial session in which they were instructed on the EMA procedure and completed demographic questionnaires and well-being measures. Initial sessions conducted during and before March 2020 took place in person (n = 118). Because of the COVID-19 pandemic, initial sessions conducted after March 2020 took place over Zoom (n = 167). Time/mode of data collection (i.e., before vs. after March 2020) did not significantly moderate any of the associations between strategy-situation fit and well-being (ps ≥ .204).

During the initial sessions, a researcher instructed participants on the EMA procedure and then provided them with a link to complete the demographic questionnaires and well-being measures. The researcher explained to the participants the importance of completing as many EMA surveys as possible and providing careful and honest responses. Community participants and half of the undergraduate participants then completed 1 week of EMA surveys (target N = 201; total analytic N postexclusions = 205), and the other half of the undergraduate participants (target N = 84; total analytic N postexclusions = 80) completed 2 weeks of EMA surveys. Participants completed EMA surveys in Smartphone Ecological Momentary Assessment (SEMA) 2 or SEMA 3 (O'Brien et al., 2024). Participants were prompted to complete EMA surveys every 72 min on average (± 30 min) between 10 a.m. and 10 p.m. for 7 days.

After cleaning the data, we observed two unexpected patterns in the iteration variable recorded by SEMA, which indicates the measurement occasion of each EMA survey (i.e., 1 for the first survey, 2 for the second survey). First, 13 participants had duplicate iteration values, making it difficult to determine the order of EMA surveys. We excluded EMA surveys with duplicated iteration values. For 10 of the 13 affected participants, only a small number of iterations were duplicated, and they all occurred within the first few surveys. Three participants had numerous duplicate iteration values throughout the EMA period, and thus these participants were excluded from analyses. Second, 42 student participants who were intended to participate for 2 weeks continued to receive and complete surveys after this 2-week period. We excluded all survey responses completed after the 140th survey for participants completing 2 weeks of surveys and all survey responses completed after the 70th survey for participants completing 1 week of surveys. In addition, a resolved SEMA bug caused responses to slider questions that were made all the way to the left or all the way to the right to be recorded outside of the range from 0 to 100. We recoded these values to 0 and 100, respectively, in line with recommendations from the SEMA team.

Measures

Well-being. Participants completed all of the well-being measures reported in the original study, including measures of depression, general anxiety, stress, neuroticism, social anxiety, and self-esteem. Table 2 shows descriptive statistics of these well-being variables in the original study and in the replication study for comparison. On average, the replication sample had lower mean well-being and greater individual differences in well-being compared with the original sample.

To assess depression, general anxiety, and stress, participants completed the 21-item Depression Anxiety Stress Scales (Henry & Crawford, 2005), a self-report

Table 2. Descriptive Statistics, Reliabilities, and Correlations Among the Well-Being Measures for the Original Study (Haines et al., 2016) and the Current Replication Study

Well-being measure	α	*M* (*SD*)	Range: Actual	Range: Possible	Correlations: 1.	2.	3.	4.	5.
Original study									
1. Depression	.86	6.38 (6.60)	0–32	0–42	—				
2. Anxiety	.77	6.05 (5.35)	0–20	0–42	.56	—			
3. Stress	.80	10.41 (7.20)	0–30	0–42	.55	.52	—		
4. Neuroticism	.86	23.70 (6.37)	9–37	8–40	.52	.55	.67	—	
5. Social anxiety	.96	81.45 (24.40)	43–141	32–160	.57	.47	.59	.76	—
6. Self-esteem	.91	31.42 (5.70)	19–40	10–40	−.58	−.41	−.38	−.60	−.66
Current replication study									
1. Depression	.91	12.06 (10.64)	0–42	0–42	—				
2. Anxiety	.84	10.12 (9.61)	0–40	0–42	.68	—			
3. Stress	.86	15.39 (10.06)	0–42	0–42	.71	.75	—		
4. Neuroticism	.85	24.91 (7.03)	8–40	8–40	.55	.51	.64	—	
5. Social anxiety	.96	88.87 (27.32)	35–156	32–160	.54	.54	.52	.62	—
6. Self-esteem	.90	28.77 (6.33)	11–40	10–40	−.62	−.44	−.51	−.70	−.57

Note: This table was adapted from the original study. Descriptive statistics are based on sum scores for the well-being measures. Scores on the Depression Anxiety Stress Scales (Henry & Crawford, 2005) were multiplied by 2 for comparison with scores on the full 42-item version of these scales. For all correlations, $N = 285$ and $p < .001$.

measure designed to assess the frequency and severity of symptoms over the course of the past week. This scale contains seven items that assess depressive symptoms, seven items that assess anxiety symptoms, and seven items that assess stress symptoms. In the original study, participants were presented with a scale that ranged from 0 (*did not apply to me at all*) to 3 (*applied to me very much, most of the time*). In the replication study, participants responded on a scale that ranged from 1 to 4, and responses were recoded to match the scale from 0 to 3 for the original study.

To assess neuroticism, participants completed the eight-item neuroticism subscale of the Big Five Inventory (John et al., 2008). Response options ranged from 1 (*disagree strongly*) to 5 (*agree strongly*).

To assess social anxiety, participants completed the 20-item Social Interaction Anxiety Scale (Mattick & Clarke, 1998) and the 12-item Brief Fear of Negative Evaluation Scale (Leary, 1983). Response options ranged from 1 (*not at all characteristic of me*) to 5 (*extremely characteristics of me*). As in the original study, the two scales were combined to create a single social-anxiety composite.

To assess self-esteem, participants completed the 10-item Rosenberg Self-Esteem Scale (Rosenberg, 1965). Response options ranged from 1 (*strongly disagree*) to 4 (*strongly agree*).

Habitual emotion regulation. In addition to the questionnaires reported in the original study, participants completed six items from the Emotion Regulation Questionnaire (ERQ; Gross & John, 2003), which assesses people's tendency to use reappraisal (α = .88). This allowed us to examine associations between strategy-situation fit and well-being above and beyond habitual reappraisal.

Impression management. In addition to the questionnaires reported in the original study, participants completed the 16-item Balanced Inventory of Desirable Responding Short Form (BIDR; Hart et al., 2015), which assesses people's tendency toward socially desirable responding. We used the impression-management subscale, which assesses bias toward pleasing others (α = .75). This allowed us to examine associations between strategy-situation fit and well-being above and beyond impression management.

EMA. In each EMA survey, participants were asked to respond to the same items that were asked in the original study. As in the original study, only a subset of these items were used to address the current research questions. Additional items assessing affect, events, and other emotion-regulation strategies were included throughout the EMA survey but were not used in analyses. In each EMA survey, participants were asked to rate the extent to which they used reappraisal "since the last survey" and the controllability of the environment "since the last survey" on a scale from 0 to 100. Reappraisal was assessed using two items: "In response to your feelings, have you looked at things from a different perspective since the

Table 3. Summary of Direct Replication and Extension Analyses

Direct replication analyses	Extension analyses
1. Associations between strategy-situation fit and six types of well-being (i.e., depression, anxiety, stress, neuroticism, social anxiety, and self-esteem), controlling for time, lagged reappraisal, and mean controllability 1a. Direction and statistical significance (α = .05) of the focal parameter 1b. Comparison of the confidence intervals around the focal parameter in the original and replication studies	1. Test–retest reliability of strategy-situation fit 2. Associations between strategy-situation fit and six types of well-being, additionally controlling for habitual reappraisal, variability in reappraisal, variability in controllability, and impression management 3. Associations between strategy-situation fit and six types of well-being in the subset of measurement occasions in which a negative or unpleasant event was experienced 4. Moderation of the associations between strategy-situation fit and six types of well-being by gender and student status 5. Associations between strategy-situation fit and six types of well-being, assessing strategy-situation fit as the association between controllability and reappraisal use (rather than controllability and change in reappraisal use)[a]

[a]This analysis was included in the supplemental materials in the original study (Haines et al., 2016).

last survey?" and "In response to your feelings, have you changed the way you were thinking about your situation since the last survey?" In the original study, the second reappraisal item referred to "the situation." We changed the item to refer to "your situation" for additional clarity. We computed a mean composite of the two reappraisal items (within-persons ω = .76; between-persons ω = .99). Controllability was assessed using the item "To what extent were you in control of what's happened since the last survey?"

At the end of the original survey, after all the items from the original study were asked, participants were also asked the binary item "Have you experienced a negative/unpleasant event since the last survey?" (0 = *no*, 1 = *yes*). Fourteen observations of the negative-event variable from four participants were collected on a scale from 0 to 100 instead of the intended binary scale and thus were recoded as missing.

Data cleaning and preparation

We treated responses made in 300 ms or less as missing, and if more than 50% of items within an EMA survey had response times less than or equal to 300 ms, the entire survey was excluded from analysis, consistent with the original study. In the original study, results were examined both including and excluding participants with low response rates (< 50%). To reduce the total number of tests and to increase reliability, we excluded participants with response rates < 50%.

Statistical analyses

All data cleaning, analyses, and visualizations were performed in R (Version 4.4.0; R Core Team, 2024) using the dplyr (Wickham et al., 2023), ggplot2 (Wickham, 2016), lubridate (Grolemund & Wickham, 2011), lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), multilevelTools (Wiley, 2020), naniar (Tierney & Cook, 2023), plyr (Wickham, 2011), psych (Revelle, 2024), readxl (Wickham & Bryan, 2023), roxygen2 (Wickham et al., 2024), stringr (Wickham, 2023), and tidyverse (Wickham et al., 2019) packages. For a summary of direct replication and extension analyses, see Table 3.

Direct replication analyses. Results from the analyses described here were used to determine the replicability of the original findings and are consistent with those used in the original study with the following exceptions. First, the original study used data from 1 week of experience sampling. A portion of our sample completed 2 weeks of experience sampling, and we used data from both weeks in primary analyses to maximize statistical power and the reliability of strategy-situation fit estimates. Second, the statistical software program Hierarchical Linear Modeling (HLM) Version 7.01 was used for all analyses in the original study. We used the lme4 package (Bates et al., 2015) in R. We chose to conduct the analyses in R because it is free, more widely used, and produces model syntax. For these reasons, it is easier to publicly share our study materials using R compared with HLM. However, we used the same estimator (restricted maximum likelihood) and the same covariance structure (unstructured) as in the original study.

Our replication analyses used models identical to those estimated in the original study. To account for nesting of observations within people, we used a series of multilevel models with random intercepts and random slopes. To test the strategy-situation fit hypothesis, we predicted reappraisal from controllability at the within-persons level while controlling for linear time. To model change in reappraisal as a function of controllability, we also controlled for reappraisal use at the previous timepoint. Lagged reappraisal for the first observation of the day was set to missing. For 1.7% of occasions, SEMA recorded the incorrect date value (e.g., 12/31/1969 for a study conducted between 2019 and 2023). Thus, we replaced all recorded dates prior to the start of data collection with missing values, and we were not able to compute lagged reappraisal for these occasions. All Level 1 variables were person-mean centered and modeled as random. The Level 1 equation is as follows:

$$\text{reappraisal}_{ti} = \pi_{0i} + \pi_{1i}\left(\text{controllability}_{ti}\right) + \pi_{2i}\left(\text{time}_{ti}\right) + \pi_{3i}\left(\text{reappraisal}_{t\text{-}1i}\right) + e_{ti} \quad (1)$$

The π_{1i} slope reflects the association between reappraisal and controllability (i.e., strategy-situation fit). At the between-persons level, well-being and mean controllability were entered as predictors and moderators of all within-persons associations. All Level 2 variables were z-scored prior to analysis. The Level 2 equations predicting the Level 1 intercept (Eq. 2) and slopes—controllability (Eq. 3), time (Eq. 4), and lagged reappraisal (Eq. 5)—are as follows

$$\pi_{0i} = \beta 00 + \beta 01\left(\text{well-being}_i\right) + \beta 02\left(\text{controllability}_i\right) + r_{0i} \quad (2)$$

$$\pi_{1i} = \beta 00 + \beta 11\left(\text{well-being}_i\right) + \beta 12\left(\text{controllability}_i\right) + r_{1i} \quad (3)$$

$$\pi_{2i} = \beta 00 + \beta 21\left(\text{well-being}_i\right) + \beta 22\left(\text{controllability}_i\right) + r_2 \quad (4)$$

$$\pi_{3i} = \beta 00 + \beta 31\left(\text{well-being}_i\right) + \beta 32\left(\text{controllability}_i\right) + r_{3i} \quad (5)$$

The key test of the strategy-situation fit hypothesis is the β11 slope. The β11 slope reflects between-persons differences in the within-persons associations of controllability and reappraisal that are associated with well-being. A separate multilevel model was tested for each of the six well-being measures (i.e., depression, anxiety, stress, neuroticism, social anxiety, and self-esteem). The original study found statistically significant ($p < .05$) β11 slopes for depression, anxiety, stress, and neuroticism and marginally significant ($p < .09$) β11 slopes for social anxiety and self-esteem.

We quantified the result of the replication attempt in two ways. First, we examined whether the focal parameter estimate (β11) was statistically significant ($p < .05$) and in the same direction as in the original study. If so, we would take this as evidence that the replication attempt was successful. Second, we examined whether the 95% confidence intervals (CIs) for the focal parameter estimate (β11) in our study overlapped with the 95% CIs in the original study. These results were used to determine whether the replication study found evidence for an effect size comparable to the effect observed in the original study. If the first criterion was met but the second criterion was not met, we would still consider the replication attempt successful, but we would note the differing effect sizes. We chose these approaches over alternative tests of replication (e.g., Bayes factor) because our approach quantifies the significance and size of the parameter estimate, whereas alternative approaches quantify its contribution to the overall model fit. This is important because we are not interested in explaining the total variance in reappraisal (the model's dependent variable) but instead in explaining variance in the within-persons association between reappraisal and controllability and its connection to well-being. Because the associations between strategy-situation fit and well-being may differ across types of well-being, results will be evaluated separately for each well-being measure. Thus, it is possible that we would find evidence for replication for some well-being measures and not others.

One potential problem that we anticipated was failure for one or more of the models to converge. We preregistered that, if this happened, we would attempt to achieve convergence by changing the optimization parameters (e.g., increasing the maximum number of iterations). If the model still did not converge, we preregistered that we would begin removing predictors in the following order: lagged reappraisal, time. These simpler models were included in the supplemental materials of the original study, and results remained largely the same. To achieve model convergence, we changed the optimization parameters, and we rescaled the EMA items by dividing by 100 to achieve model convergence. Thus, results can be interpreted on a scale from 0 to 1 rather than on a scale from 0 to 100. We chose to deviate from our preregistered plan and rescale the predictors rather than drop them because rescaling the predictors does not affect the statistical significance of the results and allowed us to stick more closely to the models used in the original study.

Extension analyses. The analyses described here were not conducted in the original study and were not used to determine the replicability of the original finding. Rather, results from these analyses were used to draw new conclusions about the strategy-situation fit hypothesis.

To estimate test–retest reliability of strategy-situation fit, individual estimates of the parameter π_{1i} (i.e., strategy-situation fit as represented by the random slopes of controllability predicting reappraisal) were extracted separately for Weeks 1 and 2 and correlated with each other in the subset of the undergraduate student sample that completed 2 weeks of EMA surveys. We assessed the distributions of the individual slope estimates for normality. Given that the reliability measures of within-person processes are typically lower than measures of mean levels, test–retest correlations greater than .30 will be considered moderate and test–retest correlations greater than .60 will be considered high.

After the Stage 1 registered report was provisionally accepted, Neubauer et al. (2020) published an alternative approach to estimate the between-persons reliability of the within-persons coupling between two variables. Thus, in addition to our planned test–retest approach, we also applied the approach described by Neubauer et al. (2020) to estimate the reliability of the strategy-situation fit parameter within the full sample.

To examine the association between strategy-situation fit and well-being above and beyond habitual reappraisal, ERQ reappraisal was entered as a Level 2 covariate in all six multilevel models. To examine the association between strategy-situation fit and well-being above and beyond variability in reappraisal use and variability in controllability, within-persons standard deviations across EMA surveys were computed for each variable. These within-persons standard deviations were included as Level 2 covariates in all six multilevel models. To rule out the possibility that good strategy-situation fit reflects participants responding to EMA items in the "right" way, we examined the association between strategy-situation fit and well-being, controlling for impression management. Specifically, BIDR impression-management scores were entered as a Level 2 covariate in all six multilevel models.

To account for the possibility that low levels of reappraisal use may be particularly prevalent in situations that are not negative or unpleasant, we assessed the association between strategy-situation fit and well-being in a subset of the data in which participants reported experiencing a negative or unpleasant event. To compare the size of the effects, we compared the CIs around the strategy-situation fit parameter in the full sample and in the subsample. We planned to interpret overlapping CIs as providing no evidence for a difference in effect size. We planned to interpret nonoverlapping CIs as providing evidence for a larger effect size in one model compared with the other.

To test for moderation by gender, we included dummy-coded gender (men compared with women) as a Level 2 moderator in all multilevel models. Specifically, we examined the interaction between gender, well-being, and controllability in predicting reappraisal use. To test for moderation by student status, we included dummy-coded student status (student participants compared with nonstudent participants) as a Level 2 moderator in all multilevel models. Specifically, we examined the interaction between student status, well-being, and controllability in predicting reappraisal use.

Finally, we examined a model without lagged reappraisal as a predictor. When lagged reappraisal is included as a predictor, as in the primary analyses in the original study, strategy-situation fit is defined as the association between controllability and change in reappraisal use. However, strategy-situation fit can also be conceptualized as the simple association between controllability and reappraisal use rather than change in reappraisal use. Thus, we examined the results of a model that is identical to the original but without lagged reappraisal as a predictor. To compare effect sizes across the two models, we planned to compare the CIs around the strategy-situation fit parameter estimate. We planned to interpret overlapping CIs as providing no evidence for a difference in effect size. We planned to interpret nonoverlapping CIs as providing evidence for a larger effect size in one model compared with the other. The original authors reported the results of these analyses in their supplemental materials, and the results were largely the same compared with the models that included lagged reappraisal.

Results

Figure 1 displays the distributions of key study variables (i.e., well-being, EMA reappraisal, and EMA controllability). The distribution of EMA reappraisal responses peaks at 0, indicating that on many occasions participants reported not using reappraisal at all. For controllability, the distribution peaks at 0 and 100, suggesting that participants experienced many situations as either completely uncontrollable or completely controllable.

Figure 2 displays time series of controllability and reappraisal for a random subsample of 25 participants. The within-persons relation between the controllability time series and the reappraisal time series is the key strategy-situation fit parameter.

Computational reproducibility

Using the data from the original study and the R code that were developed for the replication study, we were able to computationally reproduce the results from all primary analyses in the original article.

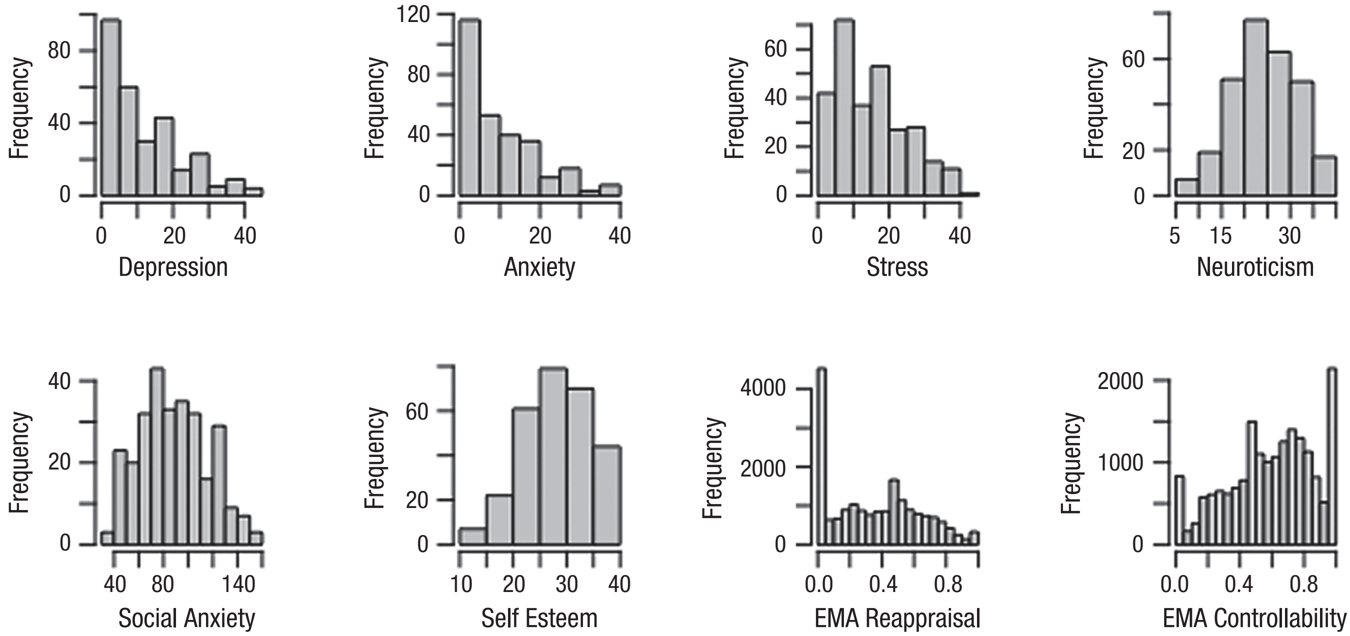

Fig. 1. Histograms of key study variables. EMA = ecological momentary assessment.

Preliminary results

Table 4 shows the means, standard deviations, and intraclass correlation coefficients for reappraisal and controllability from the original study and from the replication study. These descriptive statistics were relatively similar between the two studies.

We used a simple model with no between-persons predictors to estimate the average controllability-reappraisal slope, consistent with the original study. In the original study, the average controllability-reappraisal slope was close to zero, $\pi_{1i} = -0.06$, $SE = 0.023$, 95% CI = [−0.05, 0.04], $p = .835$. In the replication study, the average controllability-reappraisal slope was significantly positive, $\pi_{1i} = 0.084$, $SE = 0.017$, 95% CI = [0.052, 0.117], $p < .001$. Thus, for the average person, reappraisal use was positively correlated with perceived controllability. In addition, controllability-reappraisal slopes varied substantially between persons in both the original study ($SD = 0.14$) and in the replication study ($SD = 0.22$).

Direction replication analyses

Table 5 shows primary results from the original study and from the replication study. In the replication study, the focal parameter (β11) was statistically nonsignificant in all models. For all models, the 95% CI around β11 in the replication study overlapped with the 95% CI around β11 in the original study. On the basis of our preregistered decision criteria, the original study results did not replicate in this new sample. In other words, strategy-situation fit was not associated with any of the well-being measures.

Further, in the original study, average EMA reappraisal was not significantly associated with any of the well-being outcomes (β01). In the replication study, higher average EMA reappraisal was associated with significantly greater depression, anxiety, and stress and was not significantly associated with neuroticism, social anxiety, or self-esteem.

Extension analyses

Test–retest reliability. Eighty participants completed more than 70 EMA measurement occasions and were included in the test–retest reliability analyses. The test–retest data set included 3,024 observations in the first week and 2,620 observations in the second week. Test–retest reliability was .51 across 2 weeks. On the basis of our preregistered criteria, the reliability of the strategy-situation fit parameter was “moderate” and suitable for testing the strategy-situation fit hypothesis.

Reliability using the Neubauer method. Using the method described by Neubauer et al. (2020) to estimate the between-persons reliability of the within-persons coupling of reappraisal and controllability, we estimated a reliability coefficient of .80, suggesting high between-persons reliability.

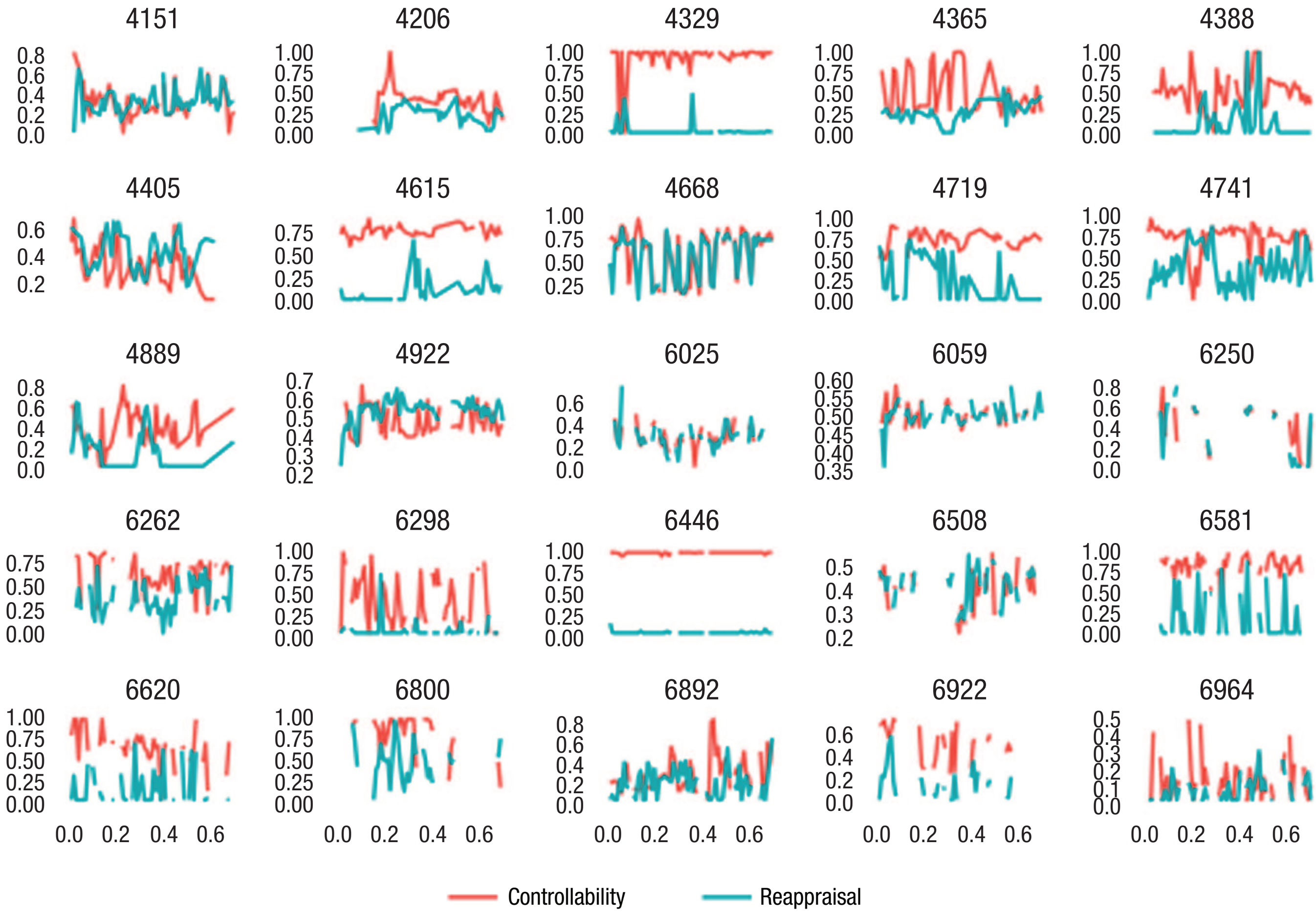

Fig. 2. Time series of controllability and reappraisal across measurement occasions for a random subsample of 25 participants.

Results adjusted for additional covariates. The focal parameter (β11) remained statistically nonsignificant in all models when individually controlling for habitual ERQ reappraisal, EMA reappraisal variability, controllability variability, and impression management (*p*s ≥ .096).

Results in negative or unpleasant contexts. After excluding measurement occasions in which no negative or unpleasant event had occurred, we were left with 2,457 observations (19% of total observations) from 256 participants (90% of total sample). When only these data were used, the focal parameter (β11) remained statistically nonsignificant in all models (*p*s ≥ .296).

Moderation by gender and student status. Gender significantly moderated the association between strategy-situation fit and neuroticism. That is, the three-way interaction between gender, neuroticism, and controllability

Table 4. Descriptive Statistics of EMA Items for the Original Study (Haines et al., 2016) and the Current Replication Study

	M	Within-persons *SD*	Between-persons *SD*	ICC
Original study				
Reappraisal	29.50	18.47	16.89	.46
Controllability	64.55	21.82	16.12	.35
Current replication study				
Reappraisal	35.95	18.52	20.06	.49
Controllability	59.85	18.99	18.16	.43

Note: The ICC indicates the proportion of the total variability in the construct that is present at the between-persons level, with the remaining variability present at the within-persons level. EMA = ecological momentary assessment; ICC = intraclass correlation coefficient.

Table 5. Fixed-Effect Estimates of Well-Being's Associations With Mean Reappraisal Use (β01) and Controllability-Reappraisal Slopes (β11) for the Original Study (Haines et al. 2016) and the Current Replication Study

	Association with mean reappraisal			Association with controllability-reappraisal slope		
Well-being measure	β01 (*SE*)	95% CI	*p*	β11 (*SE*)	95% CI	*p*
Original study						
Depression	−0.84 (1.92)	[−4.67, 2.99]	.663	0.059 (0.018)	[0.023, 0.095]	.002
Anxiety	−0.64 (1.80)	[−4.23, 2.96]	.725	0.047 (0.018)	[0.012, 0.082]	.009
Stress	3.08 (2.14)	[−1.19, 7.34]	.155	0.063 (0.017)	[0.030, 0.097]	< .001
Neuroticism	2.56 (2.19)	[−1.81, 6.93]	.246	0.050 (0.020)	[0.010, 0.089]	.014
Social anxiety	2.61 (2.02)	[−1.42, 6.63]	.201	0.035 (0.018)	[−0.001, 0.071]	.059
Self-esteem	−1.45 (2.06)	[−5.55, 2.65]	.482	−0.039 (0.022)	[−0.083, 0.006]	.088
Current replication study						
Depression	3.49 (1.22)	[1.11, 5.87]	.005	0.025 (0.016)	[−.007, .057]	.129
Anxiety	5.66 (1.18)	[3.35, 7.97]	< .001	0.019 (0.016)	[−.012, .051]	.233
Stress	4.53 (1.20)	[2.19, 6.87]	< .001	0.026 (0.016)	[−.006, .058]	.115
Neuroticism	1.89 (1.24)	[−0.53, 4.31]	.128	0.017 (0.016)	[−.015, .050]	.287
Social anxiety	1.93 (1.22)	[−0.45, 4.31]	.114	0.021 (0.016)	[−.011, .053]	.192
Self-esteem	−0.78 (1.28)	[−3.27, 1.72]	.543	−0.011 (0.017)	[−.043, .022]	.522

Note: In the current replication study, EMA variables were divided by 100 to aid in model convergence. Thus, for comparability of the original study results with the replication study results, we multiplied the β01 coefficient, its standard error, and the lower and upper bounds of its 95% CI by 100. The β11 coefficients and associated values are comparable across studies without this transformation because β11 reflects moderation of the strength of the association between two EMA variables that were on the same scale as one another in both the original and replication studies. EMA = ecological momentary assessment; CI = confidence interval.

was statistically significant (p = .047). In follow-up analyses within the subsample of men and the subsample of women, the association between strategy-situation fit and neuroticism was statistically nonsignificant and in the unexpected direction in men (β11 = −0.03, p = .231) but was statistically significant in the expected direction in women (β11 = 0.05, p = .041).

In addition, the three-way interaction between gender, self-esteem, and controllability was marginally significant (p = .065). In follow-up analyses in the subsample of men and in the subsample of women, the association between strategy-situation fit and self-esteem was statistically nonsignificant in both men (β11 = 0.02, p = .336) and women (β11 = −0.03, p = .210).

The other four three-way interactions between gender, controllability, and well-being (i.e., anxiety, social anxiety, depression, and stress) were statistically nonsignificant ($p \geq .165$). Given that the gender interaction was observed for only one of six well-being variables and in the context of 12 total moderation tests, we recommend caution when interpreting this finding.

Student status did not significantly moderate the association between strategy-situation fit and any of the six well-being variables. That is, the three-way interaction between student status, well-being, and controllability was statistically nonsignificant in all models (ps ≥ .240).

Results without adjusting for lagged reappraisal. When removing the lagged reappraisal predictor, the association between strategy-situation fit and depression was in the expected direction and was marginally significant (β11 = 0.03, p = .072). However, because the association was only marginally significant, and because this pattern was observed for only one outcome, we do not interpret this as evidence for the strategy-situation fit hypothesis. For all other well-being variables, the association between well-being and strategy-situation fit was also statistically nonsignificant (ps ≥ .217).

Exploratory analyses for habitual reappraisal. To better understand the current findings, we conducted unplanned exploratory analyses involving habitual reappraisal assessed with the ERQ. ERQ reappraisal was only modestly positively associated with participants' average EMA reappraisal, and the association was not statistically significant (r = .09, p = .118). Although the absence of a statistically significant association between average EMA reappraisal and ERQ reappraisal may seem counterintuitive, the observed effect size, r = .09, is relatively consistent with prior research that found that these two measures of reappraisal are only modestly positively correlated on average (r = .14; Koval et al., 2023).

Further, ERQ reappraisal was associated with lower depression (r = −.25, p < .001), lower anxiety (r = −.11,

$p = .057$), lower stress ($r = -.21$, $p < .001$), lower neuroticism ($r = -.37$, $p < .001$), lower social anxiety ($r = -.23$, $p < .001$), and higher self-esteem ($r = .38$, $p < .001$).

Discussion

The current preregistered direct replication study did not replicate the original finding that strategy-situation fit in daily life is associated with better well-being (Haines et al., 2016). Specifically, we did not find evidence that people who use reappraisal more during uncontrollable versus controllable situations in daily life have better well-being.

There are at least three potential reasons for the observed null results. First, the general strategy-situation fit hypothesis may be incorrect, at least as currently postulated. The strategy-situation fit hypothesis focuses on the fit between a single feature of the situation and the use of a single emotion-regulation strategy. However, multiple situational factors and aspects of the emotion-regulation process (Gross, 2015) may interact to impact well-being. Indeed, one of the key findings in research on the strategy-situation fit hypothesis is a three-way interaction between strategy ability and two features of the situation: controllability and stressor severity (Troy et al., 2013). Beyond controllability and stressor severity, other features of the situation (e.g., timing, duration, globality, threat, deprivation, event type; Epel et al., 2018) and the broader context (e.g., culture, society, community, family, and social group) may jointly impact the relationship between emotion-regulation strategy use and well-being (Mauss & Troy, 2023). Further, the fit between a particular emotion-regulation strategy and the situation may matter for well-being only when the strategy is implemented successfully. People often attempt to use emotion-regulation strategies unsuccessfully (Ford & Troy, 2019; Ford et al., 2017), which may disrupt the relationship between strategy-situation fit and well-being. Thus, the strategy-situation fit hypothesis may need to be amended to capture the multifaceted nature of both situational contexts and emotion-regulation processes.

Second, the strategy-situation fit hypothesis may be too coarse and in need of further specification. For example, it may be correct for some strategy-situation pairings (e.g., see Troy et al., 2023) but not for reappraisal and controllability. Indeed, if reappraisal is generally more adaptive in uncontrollable situations and less adaptive in controllable situations, we might expect the average person to engage in this adaptive pattern of reappraisal use. In the original study, however, reappraisal use was unrelated to controllability on average, and in this replication study, reappraisal use was positively associated with controllability on average. The adaptiveness of reappraisal in controllable contexts might also vary across reappraisal types. If an individual reappraises a controllable situation in a way that minimizes its potential consequences or changes their own goals to match the situation (Uusberg et al., 2019), this may reduce motivation to change the controllable situation and in turn may be associated with worse well-being. Alternatively, if an individual reappraises a controllable situation as challenging rather than threatening, this may increase self-efficacy to change the controllable situation and in turn may be associated with better well-being. The reappraisal items used in the current study did not distinguish between these different types of reappraisal.

Third, the current study design may not have provided an appropriate diagnostic test of the controllability-reappraisal version of the strategy-situation fit hypothesis. For example, the objective controllability of situations may be most relevant to the adaptiveness of reappraisal (e.g., Troy et al., 2013), but the current methods assessed participants' self-reported subjective ratings of controllability. Reappraisal may even influence the degree to which individuals perceive situations as controllable in a way that works against the strategy-situation fit hypothesis. An individual facing a stressful situation may try to regulate their emotions by reappraising their resources as sufficient to cope with the situation. In this example, the individual would likely report engaging in a high degree of reappraisal and high situational controllability. This would indicate "poor" strategy-situation fit according to the strategy-situation fit hypothesis despite being a potentially adaptive form of reappraisal. Reappraisal was positively associated with controllability in the current study, consistent with this possibility. The current measurement of reappraisal may have also been partially confounded with stressor severity and the accompanying need to regulate. In future research, it would be beneficial to assess strategy-situation fit in exclusively stressful or negatively valenced contexts (e.g., in the current study, participants reported a negative event on only 19% of occasions) and to statistically control for stressor severity.

The current study was limited to U.S. adults. Thus, findings may not generalize to other countries and cultures. The concept of strategy-situation fit should be relevant in other cultures; however, the objective and subjective controllability of situational contexts and emotion beliefs likely differ across sociocultural contexts.

Conclusion

In a preregistered direct replication study, we did not replicate the finding that people who use reappraisal more during uncontrollable versus controllable situations in daily life have better well-being. Potential explanations for the observed null findings highlight the complexity of associations between emotion regulation, situations, and well-being. Future empirical tests of the strategy-situation fit hypothesis and its boundary conditions will require large data-collection efforts that assess multiple aspects of the emotion-regulation process, as well as objective and subjective measures of multiple situational features.

Transparency

2018
Action Editor: Brent W. Roberts
Editor: D. Stephen Lindsay
2021
Action Editor: Patricia J. Bauer
Editor: Patricia J. Bauer
2024
Action Editor: Simine Vazire
Editor: Simine Vazire
Author Contributions

Emily C. Willroth: Conceptualization; Data curation; Formal analysis; Project administration; Writing – original draft; Writing – review & editing.

Gerald Young: Data curation; Formal analysis; Project administration; Writing – review & editing.

Brett Q. Ford: Conceptualization; Writing – review & editing.

Allison Troy: Conceptualization; Writing – review & editing.

Dorota Swierzewicz: Data curation; Formal analysis; Project administration; Writing – review & editing.

Iris B. Mauss: Conceptualization; Project administration; Resources; Supervision; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

No external funding was used for this project. Gift cards for community participants were purchased with departmental funds.

Artificial Intelligence

At the recommendation of the STAR editor team, we used ChatGPT as a resource to make some of our code more parsimonious. In all instances, we double-checked the accuracy of our code by comparing the approach ChatGPT recommended to our original approach. No other AI-assisted technologies were used in this research or the creation of this article.

Ethics

This project was approved by the institutional review board at the University of California, Berkeley (Protocol No. 2016-02-8400).

Open Practices

Study disclosures. Preregistration: This study is a registered report. The research aims, hypotheses, methods, and analysis plan were preregistered prior to data collection. The Stage 1 registered report is publicly available (https://osf.io/ap2ws). There were major and minor deviations from the preregistration (for details, see Table 1). Materials: All study materials are publicly available (https://osf.io/nk5xy). Data: De-identified data sets are publicly available (https://osf.io/u3sbr). These include observations for all participants and variables used in the analyses reported herein. Analysis scripts: All analysis scripts are publicly available (https://osf.io/gycfv). Computational reproducibility: The computational reproducibility of the results has been independently confirmed by the journal's STAR team.

ORCID iDs

Emily C. Willroth https://orcid.org/0000-0002-4780-6883
Brett Q. Ford https://orcid.org/0000-0002-7943-4447
Iris B. Mauss https://orcid.org/0000-0002-3797-4795

Acknowledgments

E. C. Willroth and G. Young contributed equally and share first/lead author position. E. C. Willroth and G. Young were affiliated with the University of California, Berkeley, during preparation of the Stage 1 registered report and during a portion of data collection, and D. Swierzewicz was affiliated with the University of California, Berkeley, during a portion of data collection and continued to contribute to the project as an independent researcher during later stages of the project. All authors gratefully acknowledge the authors of the Haines et al. (2016) study for their help, including sharing materials and code and reviewing our study procedures. We also gratefully acknowledge the contributions of our friend and collaborator Emily Butler, who coauthored the Stage 1 registered report and sadly passed away before the Stage 2 registered report was completed.

References

Aldao, A. (2013). The future of emotion regulation research capturing context. *Perspectives on Psychological Science*, *8*, 155–172.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Cheng, C., Lau, H. P. B., & Chan, M. P. S. (2014). Coping flexibility and psychological adjustment to stressful life changes: A meta-analytic review. *Psychological Bulletin*, *140*, 1582–1607.

Collins, D. L., Baum, A., & Singer, J. E. (1983). Coping with chronic stress at Three Mile Island: Psychological and biochemical evidence. *Health Psychology*, *2*, 149–166.

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, *49*, 146–169.

Folkman, S. (1984). Personal control and stress and coping processes: A theoretical analysis. *Journal of Personality and Social Psychology*, *46*, 839–852.

Ford, B. Q., Karnilowicz, H. R., & Mauss, I. B. (2017). Understanding reappraisal as a multicomponent process: The psychological health benefits of attempting to use reappraisal depend on reappraisal success. *Emotion*, *17*, 905–911.

Ford, B. Q., & Troy, A. S. (2019). Reappraisal reconsidered: A closer look at the costs of an acclaimed emotion-regulation strategy. *Current Directions in Psychological Science*, *28*, 195–203.

Green, P., & MacLeod, C. J. (2016). simr: An R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498.

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software*, *40*, 1–25.

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*, 281–291.

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*, 1–26.

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*, 348–362.

Haines, S. J., Gleeson, J., Kuppens, P., Hollenstein, T., Ciarrochi, J., Labuschagne, I., Grace, C., & Koval, P. (2016). The wisdom to know the difference: Strategy-situation fit in emotion regulation in daily life is associated with well-being. *Psychological Science*, *27*, 1651–1659. https://doi.org/10.1177/0956797616669086

Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The Balanced Inventory of Desirable Responding Short Form (BIDR-16). *SAGE Open*, *5*(4). https://doi.org/10.1177/2158244015621113

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *44*, 227–239.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). Guilford Press.

Kashdan, T. B., & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review*, *30*, 865–878.

Koval, P., Kalokerinos, E. K., Greenaway, K. H., Medland, H., Kuppens, P., Nezlek, J. B., Hinton, J. D., & Gross, J. J. (2023). Emotion regulation in everyday life: Mapping global self-reports to daily processes. *Emotion*, *23*, 357–374.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Lazarus, R. S. (1993). Coping theory and research: Past, present, and future. *Psychosomatic Medicine*, *55*, 234–247.

Leary, M. R. (1983). A brief version of the Fear of Negative Evaluation Scale. *Personality and Social Psychology Bulletin*, *9*, 371–375.

Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour Research and Therapy*, *36*, 455–470.

Mauss, I. B., & Troy, A. S. (2023). Emotion regulation and resilience: When is reappraisal (not) conducive to resilience? In J. J. Gross & B. Q. Ford (Eds.), *Handbook of emotion regulation* (3rd ed., pp. 274–281). Guilford Press.

Neubauer, A. B., Voelkle, M. C., Voss, A., & Mertens, U. K. (2020). Estimating reliability of within-person couplings in a multilevel framework. *Journal of Personality Assessment*, *102*(1), 10–21.

O'Brien, S. T., Dozo, N., Hinton, J. D. X., Moeck, E. K., Susanto, R., Jayaputera, G. T., Sinnott, R. O., Vu, D., Alvarez-Jiminez, M., Gleeson, J., & Koval, P. (2024). SEMA[3]: A free smartphone platform for daily life surveys. *Behavior Research Methods*, *56*, 7691–7706. https://doi.org/10.3758/s13428-024-02445-w

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.0) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

Revelle, W. (2024). *psych: Procedures for psychological, psychometric, and personality research* (Version 2.4.6) [Computer software]. CRAN. https://CRAN.R-project.org/package=psych

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.

Tierney, N., & Cook, D. (2023). Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software*, *105*, 1–31.

Troy, A. S., Shallcross, A. J., & Mauss, I. B. (2013). A person-by-situation approach to emotion regulation cognitive reappraisal can either help or hurt, depending on the context. *Psychological Science*, *24*, 2505–2514.

Troy, A. S., Wilhelm, F. H., Shallcross, A. J., & Mauss, I. B. (2010) Seeing the silver lining: Cognitive reappraisal ability moderates the relationship between stress and depressive symptoms. *Emotion*, *10*, 783–795.

Troy, A. S., Willroth, E. C., Shallcross, A. J., Giuliani, N. R., Gross, J. J., & Mauss, I. B. (2023). Psychological resilience: An affect-regulation framework. *Annual Review of Psychology*, *74*, 547–576.

Uusberg, A., Taxer, J. L., Yih, J., Uusberg, H., & Gross, J. J. (2019). Reappraising reappraisal. *Emotion Review*, *11*(4), 267–282.

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*, 1–29.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer.

Wickham, H. (2023). *stringr: Simple, consistent wrappers for common string operations* (Version 1.5.1) [Computer software]. CRAN. https://CRAN.R-project.org/package=stringr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2023). *readxl: Read Excel files* (Version 1.4.3) [Computer software]. CRAN. https://CRAN.R-project.org/package=readxl

Wickham, H., Danenberg, P., Csárdi, G., & Eugster, M. (2024). *roxygen2: In-line documentation for R* (Version 7.3.2) [Computer software]. CRAN. https://CRAN.R-project.org/package=roxygen2

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (Version 1.1.4) [Computer software]. CRAN. https://CRAN.R-project.org/package=dplyr

Wiley, J. (2020). *multilevelTools: Multilevel and mixed effects model diagnostics and effect sizes* (Version 0.1.1) [Computer software]. CRAN. https://CRAN.R-project.org/package=multilevelTools

Willroth, E. C., & Atherton, O. E. (2024). Best laid plans: A guide to reporting preregistration deviations. *Advances in Methods and Practices in Psychological Science*, *7*(1). https://doi.org/10.1177/25152459231213802