The Psychological Foundations of Liberal Democracy: Does Reducing Distress Increase Support for Democratic Norms and Out-Group Tolerance?*

Chagai M. Weiss† Sascha Riaz‡ Julie Cachia§ Xuan Zhao¶

November 18, 2026

Abstract

Indicators of psychological distress, including depression, anxiety, and loneliness, have risen across many industrialized countries, while support for the foundations of liberal democracy has declined. Prior research documents negative associations between distress and citizens' democratic commitments, yet causal evidence is scarce, and it remains unclear whether interventions that reduce distress have downstream political effects. We address this gap by theorizing how improved mental health may strengthen two pillars of liberal democracy—support for democratic norms and tolerance of political out-groups—and propose a seven-week randomized controlled trial that incentivizes US adults to engage with an evidence-based AI-powered well-being application shown to reduce distress. We hypothesize that app usage improves mental health and, in doing so, strengthens support for democratic norms and out-group tolerance. Our theory and evidence contribute to research on the psychological foundations of democratic citizenship, efforts to bolster liberal democracy, and debates about externalities of emerging technologies.

*This project was reviewed and approved by the University of Toronto REB. We thank the Institute for Humane Studies for generous support (grant number: IHS019684). This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant. For helpful comments and suggestions we thank Josh Kalla and Don Greene.

†University of Toronto.

‡Singapore Management University.

§Flourish Science.

¶Flourish Science.

1 Introduction

What are the effects of psychological distress on liberal democratic preferences, and specifically support for democratic norms and out-group tolerance? Over the past decade, indicators of psychological distress—including symptoms of anxiety, stress, and depression as well as persistently high levels of loneliness—have risen globally (Baarck, d'Hombres and Tintori, 2022; Daly and Macchia, 2023; Office of the Surgeon General, 2023; Liu and Kuai, 2025; Piao, Xie and Managi, 2024; Kannan and Veazie, 2023), and recent estimates suggest that over half of the world's population will experience at least one mental health condition during their lifetime (McGrath et al., 2023). Concurrently, citizens' support for the underlying foundations of liberal democracy is declining (Armingeon and Guthmann, 2014; Claassen and Magalhães, 2023; Valgarðsson et al., 2025; Druckman, 2024), while support for illiberal populists and exclusionary policies is on the rise (Mudde, 2007; Zeitzoff, 2023; Valentim, 2024; Vries and Hobolt, 2020; Valentim, Dinas and Ziblatt, 2025).

Influential policy and media accounts suggest that these global trends are closely related. In his report, "Our Epidemic of Loneliness and Isolation," the former US Surgeon General Dr. Vivek Murthy argued that loneliness is a meaningful risk factor for depression and anxiety, which can erode trust in institutions and deepen social polarization (Office of the Surgeon General, 2023). In line with this insight, various commentators suggest that psychological despair is a "threat to our health, our civil society, our democracy, and even national security" (Graham, 2024), and that rising levels of depression and anxiety lead to "more 'need for chaos' in our politics" (Thompson, 2025). These alarming statements are supported by existing research on the political consequences of psychological distress, as various symptoms of declining mental health—including loneliness, depression, anxiety, and negative affect—are associated with support for populist appeals (Peterson et al., 2025; Berlingieri, d'Hombres and Kovacic, 2025; Ward et al., 2025; Baum et al., 2025), stronger polarization (Balmas, Atia and Halperin, 2023; Edinger et al., 2025), belief in conspiracy theories (Green et al., 2023), reduced political trust (Bell et al., 2019; Chen and Shen, 2025),

decreased political efficacy and participation (Wagner and Morisi, 2019; Landwehr and Ojeda, 2021; Ojeda, Bernardi and Landwehr, 2023; Stickley et al., 2023; Young, 2019), endorsement of political violence (Baum et al., 2024), and prioritization of defensive government action at the expense of civil liberties (Albertson and Gadarian, 2015; Vasilopoulos, Marcus and Foucault, 2018; Zeitzoff, 2014; García-Ponce, Young and Zeitzoff, 2023).[1]

Yet previous studies examining the effects of psychological distress on support for the foundations of liberal democracy—which we conceptualize as support for democratic norms (Levitsky and Ziblatt, 2019) and tolerance of political out-groups (Sullivan, Piereson and Marcus, 1993)—face three major limitations. The first is causal identification.[2] There are sound theoretical reasons to hypothesize that psychological distress may erode support for liberal democracy. But distress is a product of complex and interrelated individual attributes and social processes—including personality traits, social esteem, economic vulnerability, and adverse life experiences (Shri, 2010)—which are known to have independent effects on political outcomes. Accordingly, when discussing the emerging literature on the political consequences of mental health, James Druckman emphasizes that "it's very difficult to think about causal studies of depression," and overcoming identification challenges would likely require randomized controlled trials, which are especially complex to design and implement (Institute for Policy Research, 2023).

The second challenge facing this literature relates to measurement and has important theoretical implications. Existing studies often focus on specific aspects of psychological distress (e.g., depression, stress, anger, anxiety, or loneliness), and rarely report the correlates of political outcomes with a broad range of clinically validated psychological metrics. Thus, it remains unclear whether different elements of distress have diverging (or similar) social and political consequences. Addressing this limitation requires conceptualizing psychological distress as a broad and multidimensional phenomenon, and using state-of-the-art psychological and clinical measures to capture its different manifestations, including depressive symptoms (Kroenke et al., 2009), anxiety symp-

[1] Notably, other research suggests that social esteem may have opposite effects in the domain of political participation (McClendon, 2014).

[2] See, for example, Ladd and Lenz (2008) on the challenges of evaluating the effects of anxiety on vote choice.

toms (Spitzer et al., 2006), stress (Cohen, Kamarck and Mermelstein, 1983), loneliness (Hughes et al., 2004), affective experience (Jenkins et al., 2023), and flourishing (Diener et al., 2010).

Finally, from an applied perspective, studies on psychological distress and support for democracy document descriptive patterns with important policy implications. One reading of this literature is that declining mental health might erode public support for the foundations of liberal democracy, and developing policies and tools to reduce distress may be a productive way to promote democratic resilience (Institute for Policy Research, 2023). Yet despite a large literature testing applied interventions to reduce polarization and increase support for democracy (Hartman et al., 2022; Voelkel et al., 2024; Weiss, Green and Willer, Forthcoming), mental health support has not, to our knowledge, been rigorously evaluated as an antidote to democratic erosion.

To address these limitations and examine the effects of psychological distress on support for the foundations of liberal democracy, we collaborate with *Flourish*: a purpose-built, AI-powered well-being application that integrates evidence-based techniques from cognitive-behavioral therapy (CBT), dialectical-behavioral therapy (DBT), acceptance-and-commitment therapy (ACT), and positive psychology. Specifically, we design a randomized controlled trial (RCT) in which we incentivize US participants to use the *Flourish* app regularly for six weeks, and then elicit social and political outcomes through a seemingly unrelated survey. Recent RCTs find that using the *Flourish* app over six weeks improves psychological well-being and reduces loneliness, anxiety, and depression by roughly 0.2–0.3 SD across mental-health indices (Cachia et al., 2026a,b), consistent with a broader literature that documents the efficacy of AI-delivered CBT for reducing depression and anxiety (Zhong, Luo and Zhang, 2024; Limpanopparat, Gibson and Harris, 2024; Heinz et al., 2025; Angelucci, Fabregas and Vazquez, 2026). Drawing on this evidence, the primary objective of our study is to examine whether encouragement to use the *Flourish* app affects mental health outcomes and, thereby, has downstream effects on support for the foundations of liberal democracy.

1.1 Suitability for registered report format

This proposed RCT is especially well-suited to a registered report for three primary reasons. First, our research design is well-positioned to examine the relationship between psychological distress and support for the underlying foundations of liberal democracy. Specifically, our empirical paradigm—an encouragement RCT design—has been used successfully to study a range of complex topics in political science, including the effects of affective polarization (Broockman, Kalla and Westwood, 2023), consuming cross-cutting news (Broockman and Kalla, 2025), and using social media (Asimovic et al., 2021). Moreover, our treatment—the use of the *Flourish* app—has been shown to yield substantial improvements in mental health, even when compliance with treatment was neither forced nor incentivized (Cachia et al., 2026a,b). Notably, our proposed design will provide strong incentives to maximize compliance, increase the magnitude of treatment effects on mental health, and allow a credible test of the downstream effects of mental health on democratic attitudes.

Second, any plausible pattern of results from our RCT will provide novel and theoretically important empirical evidence to the literature on the psychological foundations of democratic citizenship. Specifically, rejecting all our null hypotheses would provide strong causal evidence supporting theories of the relationship between psychological distress and support for the foundations of liberal democracy. Alternatively, rejecting our first null hypothesis relating to mental health without rejecting the null hypotheses relating to downstream political outcomes would suggest that correlational estimates in the literature should be interpreted with caution and that the link between psychological distress and support for liberal democracy is unlikely to be causal, as proposed by existing theory and evidence. Though previous RCTs (Cachia et al., 2026a; Zhong, Luo and Zhang, 2024; Limpanopparat, Gibson and Harris, 2024; Heinz et al., 2025), as well as our highly powered design, render this outcome unlikely, failing to reject the null for both mental health and political hypotheses would emphasize that emerging technologies may be ineffective at meaningfully addressing the national mental health crisis, despite scholarly and popular attention

to their promising potential. In turn, to the extent that declining mental health has adverse political consequences, our findings would imply that undoing those consequences would require other policies and interventions.

Finally, to date, most empirical investigations into the political consequences of declining mental health and psychological despair have reported positive results (Peterson et al., 2025; Edinger et al., 2025; Chen and Shen, 2025; Ojeda, Bernardi and Landwehr, 2023; Stickley et al., 2023; Baum et al., 2024).[3] These results may reflect the true distribution of empirical estimates in the field. But a skeptic might worry that publication bias (or threats to causal identification) could skew the distribution of estimates (Franco, Malhotra and Simonovits, 2014; Moniz, Druckman and Freese, 2025), thereby adversely influencing our broader understanding of the political consequences of psychological distress. A registered report maximizes the credibility of our research program and ensures that empirical estimates of the political effects of declining mental health are broadly shared, regardless of their alignment with evidence from previous correlational studies.

1.2 Main results and contributions

Through our proposed RCT, we will identify how access to an evidence-based emerging technology designed to improve mental health shapes participants' psychological distress, whether app-induced reductions in distress shift intermediate psychological orientations—including zero-sum thinking (Chinoy et al., 2026), conversational receptiveness (Yeomans et al., 2020), political anxiety (Weinschenk et al., 2026), and pessimism (Scheier, Carver and Bridges, 1994)—and whether these changes have downstream consequences for two families of political outcomes. The first concerns support for liberal-democratic norms: (i) support for in-party norm violations, (ii) opposition to political violence, and (iii) support for bipartisan cooperation (Voelkel et al., 2024; Holliday et al., 2024; Weiss, Green and Willer, Forthcoming). The second concerns attitudes toward various out-groups: (i) generalized trust, (ii) a partisan out-group thermometer, and (iii) affective polariza-

[3] One exception is Haushofer et al. (2023), who experimentally increase levels of the stress hormone cortisol among Kenyan lab participants and find limited support for the relationship between stress, prosociality, and coethnic preferences.

tion (Holliday et al., 2024; Jha, Shayo and Weiss, 2025). As secondary outcomes, we also examine political interest and turnout intention (Ojeda, Bernardi and Landwehr, 2023), as well as respondents' beliefs about the out-party's compliance with democratic norms and its support for political violence.

Our study will contribute to three lines of research. First, we contribute to an emerging literature on the psychological foundations of support for liberal democracy. Complementing recent studies that report associations between various elements of psychological distress and adverse political outcomes (Baum et al., 2024; Albertson and Gadarian, 2015; Baum et al., 2025; Balmas, Atia and Halperin, 2023; Green et al., 2023), we provide novel causal estimates on the effects of (improved) mental health on support for the foundations of liberal democracy.

Second, we contribute to a broad literature that develops and tests approaches to reduce polarization and increase support for democracy (Voelkel et al., 2024; Weiss, Green and Willer, Forthcoming; Hartman et al., 2022; Druckman et al., 2023). This literature focuses overwhelmingly on information provision and psychologically informed messaging, often implemented via light-touch interventions which are evaluated in highly controlled survey experiments. Skeptics emphasize that the applied impact of these approaches may be minimal, due to their small and fleeting effects (Holliday, Lelkes and Westwood, 2025), and the fact that in the absence of forced exposure, most citizens lack the motivation or incentive to constructively engage with counter-attitudinal information or experiences that increase support for democracy (Landry and Halperin, 2025; Weiss, Siegel and Scacco, Forthcoming). Against this backdrop, we evaluate a non-paternalistic and incentive-compatible intervention that may increase support for democracy. Our subjects' engagement with the *Flourish* app is designed to improve their *personal* well-being. A potential positive externality of this welfare-enhancing engagement may be increased support for democracy. Notably, unlike other approaches that require substantial resources to deliver at-scale change, the intervention we study can be scaled at near-zero marginal cost, especially among populations lacking access to traditional mental health support.[4]

[4]Though we wish to emphasize that our intervention, is best thought of as a complement, rather than a substitute to traditional mental health support.

Finally, we contribute to research on the political consequences of emerging technologies. Political scientists are increasingly examining citizens' preferences over AI policy and regulation (Magistro et al., 2026; Borwein et al., 2026; Westwood, Grimmer and Hall, 2025; Zilinsky and Zeitzoff, 2024). Yet we know less about the broader social and political externalities of using AI-powered technologies. We focus on mental health, a domain in which human-delivered care faces a hard scalability ceiling, due in part to chronic workforce shortages (Clement et al., 2015; Kazdin and Blase, 2011; Dubois, 2025). Specifically, we examine the consequences of engaging with *AI-powered* mental health support, which is becoming increasingly common because of its potential to complement human care—especially for individuals in underserved communities—by providing private, secure, round-the-clock support at near-zero marginal cost. Whether this technological development has downstream consequences for democratic health is the question we explore in the pages that follow.

2 Theoretical framework and hypotheses

Our primary goal is to evaluate the relationship between psychological distress and support for the underlying foundations of liberal democracy, and consider the extent to which evidence-based technologies designed to improve mental health have downstream political effects. To do so, we develop a theoretical framework that follows two steps. First, we briefly explain why using an evidence-based mental-health support application may meaningfully affect psychological distress and broader well-being. Then, turning to the primary component of our theoretical framework, we explain why reduced psychological distress and improved mental health may have important political implications.

2.1 Upstream effects: Mental health support reduces psychological distress

Digital mental-health interventions are a scalable way to expand access to psychological support, especially for people who face barriers to traditional care. App- and web-based programs, espe-

cially those grounded in cognitive-behavioral principles, have demonstrated efficacy in reducing depression and anxiety symptoms (Linardon et al., 2019; Andersson et al., 2019). More recently, AI- and chatbot-delivered interventions have shown promise in delivering structured psychological support at scale, with emerging randomized (Heinz et al., 2025) and meta-analytic (Zhong, Luo and Zhang, 2024) evidence indicating reductions in depression, anxiety, and related forms of distress.

The *Flourish* application, which we employ in our study, flexibly draws from multiple evidence-based techniques to tailor support to users' unique circumstances. It delivers personalized well-being support through *Sunnie*, an AI well-being coach built on a purpose-built, safety-first architecture that integrates psychological science, longitudinal personalization, and real-time risk detection, alongside interactive journaling, brief science-based activities, and habit-building. The platform integrates techniques from cognitive-behavioral therapy, which targets maladaptive cognitions and behaviors (Beck and Alford, 2014); dialectical-behavioral therapy, which emphasizes emotion regulation and distress tolerance (Linehan and Wilks, 2015); acceptance and commitment therapy, which focuses on psychological flexibility, acceptance, and values-based action (Hayes et al., 2006); and positive psychology, which aims to build psychological resources such as positive emotion, meaning, gratitude, and social connection (Sin and Lyubomirsky, 2009; Bolier et al., 2013).

Prior RCTs of the *Flourish* app provide evidence that access to the platform improves multiple dimensions of mental health and well-being. Specifically, six weeks of app use have been shown to cause meaningful reductions in anxiety and depression (Cachia et al., 2026b), and to increase positive affect, resilience, and social well-being (Cachia et al., 2026a). We therefore expect that encouraging study participants to use the *Flourish* app will reduce psychological distress and improve multiple aspects of study participants' mental health.

2.2 Downstream effects: Reductions in psychological distress shape political outcomes

Drawing on existing research showing that various elements of psychological distress—including depression (Baum et al., 2024), anxiety (Albertson and Gadarian, 2015), stress (Hirsch-Hoefler et al., 2016), and loneliness (Peterson et al., 2025)—are politically consequential, we theorize that app-induced improvements in mental health may increase support for the underlying foundations of liberal democracy. Whereas past studies typically examine the relationship between specific manifestations of psychological distress (e.g., depression or anxiety) and particular political outcomes (e.g., support for political violence or belief in conspiracy theories), our theoretical framework is broader. We examine how changes in mental health—broadly defined as a multidimensional construct encompassing both psychological distress and positive psychological well-being—shape a range of political attitudes and preferences that undergird liberal democracy. But why might mental health interventions affect people's support for the core foundations of liberal democracy?

We theorize that mental health support provides people with practices and skills that can help them manage stress and anxiety, regulate negative affect, and reflect on the connections between their emotions, beliefs, preferences, and behaviors (Beck, 1970; Hayes et al., 2006; Kliem, Kröger and Kosfelder, 2010; Heller et al., 2017; Blattman, Jamison and Sheridan, 2017; Blattman et al., 2023). Although these practices are usually expected to influence personal growth and psychological well-being, we argue that the skills and capacities acquired through mental health interventions may also influence four politically relevant psychological dispositions: political anxiety, pessimism, zero-sum thinking, and openness to others. By influencing these dispositions, mental health interventions may ultimately increase support for the underlying foundations of liberal democracy. We elaborate on this below and summarize our argument in Table 1.

Political anxiety. Anxiety is a core dimension of mental health and a correlate of hostile intergroup attitudes, support for restrictive security policies, and isolationist preferences (Balmas,

Harel and Halperin, 2022; Albertson and Gadarian, 2015; Vasilopoulos, Marcus and Foucault, 2018; Huddy, Feldman and Weber, 2007). Anxious individuals are more likely to perceive uncertainty as danger and to overestimate the likelihood and severity of negative outcomes in ambiguous contexts. Mental health interventions are expected to reduce anxiety, and in turn help individuals reappraise threatening situations and reduce reliance on automatic, threat-focused interpretations of their social context.

As noted above, we expect engagement with mental health interventions to reduce general anxiety. Recent research, however, distinguishes between general anxiety and *political* anxiety (Weinschenk and Smith, 2024). Whereas the former reflects a broad psychological disposition, the latter captures anxious reactions to political events, actors, and experiences. By reducing both general and political anxiety, mental health interventions may shape how individuals engage with politics. Less anxious individuals may approach political contestation more level-headed and be slower to read uncertainty as danger, compromise as weakness, or opponents as threats. That, in turn, should make them more willing to tolerate ambiguity, to weigh views they do not share, and to accept the restraint, pluralism, and peaceful contestation that democratic practice demands. In this way, reductions in anxiety may increase support for democratic norms (specifically, opposition to political violence), and improve intergroup attitudes (specifically, generalized trust and affect toward out-groups).

Pessimism. Relatedly, mental health interventions that reduce anxiety and improve well-being may shape people's outlook on the world, encouraging more optimistic rather than pessimistic worldviews. Such shifts may be especially meaningful for politics. Existing research conceptualizes pessimism as the belief that society is in decline and that improving collective outcomes is beyond society's capacity (Steenvoorden and Harteveld, 2018).

In line with this understanding, pessimism is theorized to demobilize citizens (Ojeda, Bernardi and Landwehr, 2023) and has been shown to increase support for populist politicians (Steenvoorden and Harteveld, 2018). We therefore expect pessimistic citizens to see democratic norms

as fragile and easily violated, to doubt that compromise pays, that opponents will reciprocate restraint, and that others can be trusted in general. If improvements in mental health reduce pessimism and foster greater optimism, they may strengthen citizens' confidence that democratic engagement, compromise, mutual restraint, and trust can produce meaningful collective gains. Thus, optimism may bolster support for democratic norms (specifically, bipartisan cooperation and rejection of outparty norm violation), and increase outgroup tolerance (specifically, generalized trust), as optimistic citizens are more likely to believe that out-partisans, and others more broadly, will reciprocate restraint and join them in pursuing the common good.

Zero-sum thinking. Zero-sum thinking refers to the belief that social and political life is fundamentally rivalrous, and that gains for one group necessarily come at the cost of another (Chinoy et al., 2026). Past studies show that such beliefs are closely related to psychological distress. Specifically, loneliness, low life satisfaction, and negative affect have all been shown to correlate with perceptions of social environments being zero-sum rather than positive-sum (Borawski, 2018; Różycka-Tran et al., 2021).

Importantly, existing evidence links zero-sum thinking to the avoidance of cross-partisan political conversations, resistance to engaging with opposing views, and heightened adversarial orientations toward political opponents (Andrews Fearon and Götz, 2024; Stefaniak, Mallett and Wohl, 2020). Mental health interventions, however, are specifically designed to challenge maladaptive cognitive patterns, including rigid zero-sum interpretations of social interaction. If they help people see positive-sum possibilities around them, and specifically realize that compromise and cooperation can produce mutual gains, improvements in mental health may strengthen support for the underlying foundations of liberal democracy. Thus, we expect that reduced zero-sum thinking will affect both support for democratic norms (specifically, support for bipartisan cooperation) and out-group attitudes (specifically, affective polarization and generalized trust).

Openness to others. Finally, improvements to mental health may shape support for the underlying foundations of liberal democracy by influencing individuals' openness toward their out-groups

and toward counter-attitudinal perspectives more broadly (Yeomans et al., 2020; Minson and Chen, 2022). Existing research suggests that psychological distress narrows social attention and heightens threat perceptions (Bar-Haim et al., 2007; Britton et al., 2011). In turn, distressed individuals are more likely to avoid their out-groups and adopt hostile attitudes toward them (Balmas, Harel and Halperin, 2022; Huddy, Feldman and Weber, 2007; Arceneaux, 2017).

Avoidance of out-groups and their perspectives may be politically consequential (Weiss, Siegel and Scacco, Forthcoming; Weiss, 2026). Citizens who are closed to cross-partisan engagement are more prone to polarization (Minson and Chen, 2022). In turn, they may be less willing to extend generalized trust and more likely to perceive political opponents as bad-faith actors rather than legitimate democratic competitors (Lee, 2022; Iyengar and Westwood, 2015). Mental health interventions—particularly those grounded in acceptance and commitment therapy and positive psychology—are designed to cultivate non-judgmental awareness, psychological flexibility, and openness (Hayes et al., 2006; Sin and Lyubomirsky, 2009; Bolier et al., 2013). By increasing receptiveness to out-groups and their perspectives, improvements in mental health may strengthen citizens' willingness to engage in compromise and cooperation with out-groups toward the common good—an important component of our support for democratic norms index—and also increase generalized trust and affect toward the out-group, while reducing affective polarization.

2.2.1 Theoretical expectations

Our primary expectation is that improvements to mental health have downstream political consequences, ultimately increasing support for the underlying foundations of liberal democracy. Drawing on our theoretical framework above, and as summarized in Table 1, improved mental health may do so by activating four politically relevant psychological mechanisms: reduced political anxiety, lower pessimism, diminished zero-sum thinking, and increased openness to others. Importantly, these mechanisms are neither independent nor mutually exclusive. Our primary objective, however, is to leverage the *Flourish* app to estimate the general downstream effects of mental health on support for liberal democracy, rather than to evaluate the app and its most effective com-

ponents, or to identify the specific pathways through which improvements in mental health shape political outcomes. Indeed, we argue that these four mechanisms may be closely related. Reductions in anxiety may attenuate pessimism. Reduced pessimism, in turn, may reduce zero-sum cognition, and reduced zero-sum cognition may encourage greater openness to competing perspectives. This highlights the potential complementarity of these mechanisms, as well as the possibility that mental health interventions may influence multiple psychological dispositions that reinforce one another. Because the *Flourish* app itself bundles brief activities that draw on multiple therapeutic approaches (Section 2.1), our design likewise does not allow us to attribute downstream effects to any single component of the app.

Thus, mental health interventions that address these underlying tendencies may generate compounding improvements across multiple dimensions simultaneously. Insofar as liberal democracy requires citizens to tolerate uncertainty, extend trust across partisan lines, regard compromise as worthwhile, perceive growth and improvement as realistic, and accept the legitimacy of political competition, improvements in mental health may cultivate the psychological foundations on which a healthy democratic culture depends. Accordingly, in light of this theoretical framework, we test four hypotheses, ordered to follow the causal chain from mental health to intermediate psychological orientations to political outcomes:

H1. Access to AI-based mental health support improves mental health and well-being.

H2. Access to AI-based mental health support shifts intermediate psychological orientations—reducing zero-sum thinking and political anxiety, increasing conversational receptiveness, and reducing pessimism.

H3a. Access to AI-based mental health support strengthens support for liberal-democratic norms.

H3b. Access to AI-based mental health support improves attitudes toward out-groups.

Table 1: Hypothesized links between intermediate psychological orientations (H2) and the two families of political outcomes: support for democratic norms (H3a) and out-group attitudes (H3b).

Psychological construct	Political outcome(s)	Theorized mechanism
Panel A. Support for democratic norms (H3a)		
Political anxiety (↓)	Opposition to political violence	Reductions in anxiety reduce threat-driven reasoning, making citizens less likely to endorse aggressive responses, and more likely to support restraint and peaceful contestation.
Pessimism (↓)	Bipartisan cooperation; rejection of in-party norm violations	Restoring optimism renews confidence that cooperation and restraint can yield collective gains, so democratic norms no longer appear fragile and citizens are less inclined to excuse their own party's violations.
Zero-sum thinking (↓)	Bipartisan cooperation	Recognizing positive-sum possibilities emphasizes the benefits of cooperation.
Openness to others (↑)	Bipartisan cooperation	Greater willingness to engage opposing perspectives raises readiness to compromise.
Panel B. Attitudes toward out-groups (H3b)		
Political anxiety (↓)	Out-group thermometer; affective polarization	Citizens with lower anxiety are less likely to view out-partisans as threatening, leading to warmer evaluations of the out-group.
Pessimism (↓)	Generalized trust	Restoring optimism renews confidence that others will reciprocate cooperation, leading to higher generalized trust.
Zero-sum thinking (↓)	Affective polarization; generalized trust	Abandoning the belief that out-party gains are one's own losses lowers adversarial orientations toward political opponents and increases general trust in others.
Openness to others (↑)	Generalized trust; out-group thermometer; affective polarization	Openness increases engagement with counter-attitudinal perspectives and political opponents, leading to more positive evaluations of out-groups and extended generalized trust.

Note: Outcomes are organized by the two downstream families in Table 2. Each row lists the outcome(s) an orientation is expected to move most proximally.

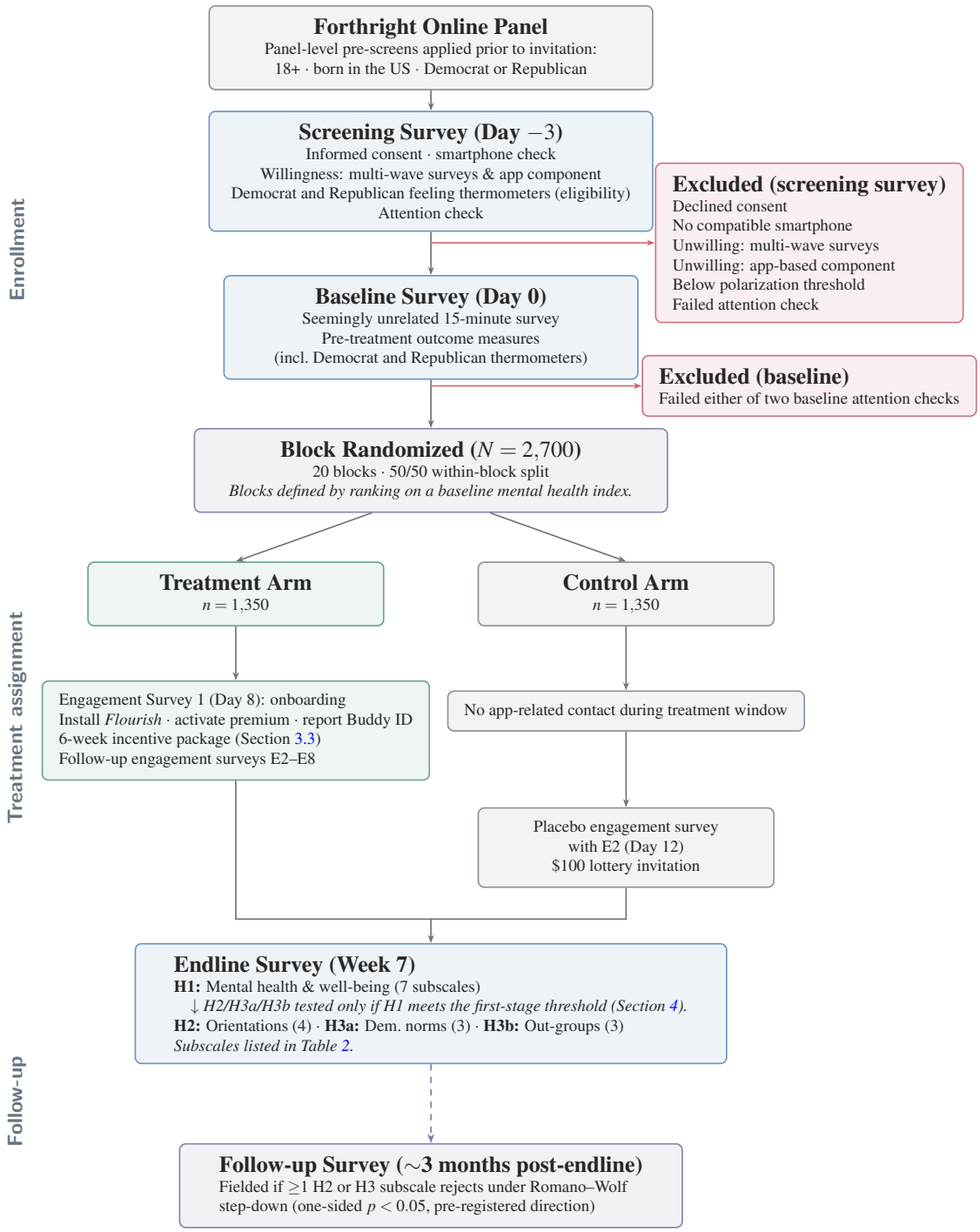
3 Research design

We test our main hypotheses in a pre-registered randomized controlled trial of $N = 2{,}700$ US adults, fielded over a seven-week window. Treated respondents receive three months of free premium access to *Flourish*, an AI-powered well-being application, paired with an incentive architecture that sustains engagement (Section 3.2); control respondents receive no app access and no app-related contact.[5] We test four pre-registered families of outcomes—mental health and well-being (H1), intermediate psychological orientations (H2), democratic norms (H3a), and out-group attitudes (H3b)—alongside a small set of exploratory secondary outcomes (Section 3.4). The Research Ethics Board at the first author's institution has reviewed and approved the full protocol (Section 3.5). Below, we provide more details on all aspects of the research design (visualized in Figure 1).

To test the duration of treatment effects, we plan to field a follow-up survey approximately three months after the endline. The follow-up uses the same instrument as the endline survey (SI Appendix A.2). We field the follow-up only if at least one subscale in the H2 (intermediate psychological orientations) or H3 (political outcomes) families has a Romano–Wolf-adjusted p-value below 0.05 in the pre-registered direction at endline. This decision is motivated by our expectation that improvements in intermediate psychological outcomes may translate into downstream H3 effects over a horizon longer than the six-week treatment window. We will estimate the same ITT specification (Equation 1) on the follow-up wave.

[5] We choose to employ an empty control rather than a placebo because other, similar app-based treatments may themselves affect our primary outcomes. A non-purpose-built chatbot, for instance, may impair mental health (Fréget, Reshef and Senik, 2026), while an app like Duolingo may expose respondents to other cultures and thereby improve intergroup attitudes (Weiss, Ran and Halperin, 2023). Less-intensive apps that we would expect to leave our primary outcomes unchanged (e.g., a weather app) are not suitable placebos either, because they demand far less engagement than the Flourish app. This choice implies that our treatment is a bundle comprising (i) the Flourish app, (ii) the reminders to use it, and (iii) the incentive package delivered after the endline survey. We have little reason to expect the reminders themselves to affect our outcomes. To minimize incentive-based confounding, however, we take two steps. First, we delay all incentive deliveries until after the endline survey. Second, through a placebo engagement survey, we inform control-group subjects that they are invited to a lottery of five $100 prizes, drawn once after the endline survey, as a token of appreciation for their participation.

Figure 1: Study flow diagram.



Note: Dashed arrow marks the conditional follow-up wave, fielded only if at least one H2 or H3 subscale rejects under the Romano–Wolf step-down at endline (Section 3). All control-arm respondents receive a placebo engagement survey, fielded alongside E2 (Day 12), inviting them to a lottery of five $100 prizes, drawn once at a predetermined date after the endline survey (SI Appendix A.6).

3.1 Sample, eligibility, and randomization

We recruit 2,700 eligible respondents through Forthright, a US-based online panel provider.[6] We establish eligibility through a combination of panel-level pre-screens applied prior to survey invitation and a standalone screening survey fielded three days before baseline (day -3; full wording in SI Appendix A.1). Respondents must (i) be at least 18 years old and born in the US; (ii) complete the survey on an iOS or Android smartphone capable of installing the *Flourish* application; (iii) express willingness to participate in multiple survey waves and in a six-week app-based component of the study (screener items Q1 and Q2; SI Appendix A.1); (iv) identify as a Democrat or a Republican; (v) display affective polarization (in-party minus out-party feeling thermometer, measured in the screening survey) above the 25th percentile of the corresponding distribution among partisans in the most recent ANES survey wave available at the time of the screening survey;[7] and (vi) pass an instructed-response attention check embedded in the screening survey, as well as two embedded attention checks in the baseline survey instrument (SI Appendix A.3).

Three days after screening, eligible respondents are invited to complete a 15-minute baseline survey that records pre-treatment measures of mental health, well-being, and political outcomes. To minimize demand effects, this baseline survey does not mention the original screener, the Flourish app, or our broader incentive package. Complementing our baseline survey, Forthright additionally supplies a set of pre-treatment covariates from the panel profile for every respondent, including age, US-born, gender, and partisanship; age, US-born, and partisanship enter the eligibility screens (above), gender and partisanship define pre-registered subgroup CATEs (Section 4.4), and none are re-asked in the survey.

Conditional on completion of the baseline survey, we randomize treatment assignment within blocks defined by baseline mental health. We first construct a single baseline mental-health index

[6] Forthright has been used to implement similar designs (Weiss, Green and Willer, Forthcoming; Hanlon et al., 2026), due to its high retention rates that exceed 90% over a seven-week study period.

[7] We retain respondents in the upper three quartiles of the partisan polarization distribution. The exact threshold value will be inserted prior to fielding based on the latest ANES release (see SI Appendix A.1). The Democrat and Republican feeling thermometers are administered in the screening survey for this eligibility check and re-administered at baseline and endline alongside the rest of the H3b feeling-thermometer battery (Section 3.4); the baseline-wave measurements serve as the pre-treatment values for the affective-polarization subscale.

for every eligible respondent: the equal-weighted mean of the seven z-standardized H1 subscales in Table 2, with the five distress subscales reverse-coded so that higher index values correspond to better mental health. We rank eligible respondents on this index and divide them into 20 equal-sized blocks. Within each block we randomly assign half of the respondents to the treatment arm and half to a control arm, so that $P(\text{treat} \mid b) \approx 0.5$ in every block b.[8] We use this index only for randomization blocking and conduct outcome tests at the subscale level (Section 3.4).

We separate the screener, baseline, onboarding, and endline surveys in time, framing, and visual display. The screener (implemented on Day -3 of our study), which includes our consent form and screening-necessary survey items, tells respondents that they are invited to participate in a survey-based study on mental health, well-being, and perspectives on public life, and that some of them may be invited to a module involving a well-being mobile app. Notably, at this stage, and throughout all baseline and endline surveys, we do not mention the Flourish app, the incentive package, or a respondent's treatment assignment.[9] On Day 0, we use a seemingly unrelated survey to elicit all baseline survey measures. Finally, on Day 8 (eight days after baseline), we invite only respondents assigned to treatment, through the first engagement survey (E1, which doubles as the onboarding survey), to download the app (SI Appendix E) and enter the incentive architecture (see Section 3.3). We do not contact control respondents about the app. However, we invite control-group respondents to participate in a brief placebo engagement survey (fielded at the same time as E2, on Day 12) where they are informed that they will be added to a lottery in which five \$100 prizes are awarded in a single draw at a predetermined date after the endline survey (see SI Appendix A.6). The purpose of this placebo is to minimize differences in incentive expectations across the treatment and control groups. After the treatment period, we recontact all respondents for an endline survey.

Previous studies using similar designs—in which outcome surveys are presented as unrelated

[8] If a block contains an odd number of respondents (n_b), we assign $\lfloor n_b/2 \rfloor$ respondents to one arm and $\lceil n_b/2 \rceil$ to the other, alternating the arm that receives the extra respondent across blocks so that the two arms remain balanced in aggregate.

[9] Two items on treated respondents' experience with the Flourish app are asked in the last engagement survey (E8; SI Appendix A.5.1).

to treatment (Jha and Shayo, 2019; Jha, Shayo and Weiss, 2025), including on the Forthright panel (Hanlon et al., 2026; Weiss, Green and Willer, Forthcoming)—do not encounter differential attrition. Supplementary analyses in these studies further show that an overwhelming majority of respondents neither connect treatment delivery to outcome measurement nor recognize the study's primary objective, easing concerns about experimenter demand effects (Jha and Shayo, 2019; Hanlon et al., 2026). Such effects, following Blair, Coppock and Moor (2020), arise when respondents seek to satisfy a social referent—here, the researcher—who they believe will observe their responses, and when doing so carries some consequence. Our design weakens all three conditions: by disassociating treatments from outcomes and never disclosing our hypotheses, respondents are unlikely to infer our expectations or to believe their answers are visible to us; and because compensation is fixed and respondents have no relational contact with our team, confirming any hypothesis offers no social reward. We nonetheless take four additional steps to further minimize such concerns.

First, to minimize attrition at endline, we will send routine reminders via Forthright after the initial rollout of the endline survey, and recontact non-responders (from both the treatment and control groups) after a week, doubling the original survey incentive. In a recent study implemented with Forthright, this approach yielded a 94% (78%) response rate six weeks (eight months) post-treatment, with no differential attrition across experimental conditions (Hanlon et al., 2026). Second, as described further in Section 4, we will formally test for differential attrition[10] and complement our main estimates with estimates using sharp Lee bounds (Lee, 2009) if we reject the null hypothesis that treatment does not cause differential attrition ($p < 0.05$, two-tailed test). Third, to further disassociate our seemingly unrelated surveys, we ensure that the screener, baseline, treatment, and endline surveys all use different visual templates. Finally, after all outcome measures in the endline survey, we elicit respondents' beliefs about our study's main objective with an open-ended question and will report the number of respondents who correctly identify it. We will use responses to this question in exploratory analyses reported in the appendix to gauge the scope of

[10]Defined as no reporting of outcomes at endline.

potential demand effects (which we expect to be minimal).[11]

3.2 Intervention and usage measurement

We deliver the treatment through a partnership with *Flourish*, an AI-powered well-being application that integrates techniques from cognitive-behavioral therapy (CBT), dialectical-behavioral therapy (DBT), acceptance-and-commitment therapy (ACT), and positive psychology. Recent RCTs find that using the *Flourish* app over six weeks meaningfully improves psychological well-being across mental-health indices (0.2–0.3 SD) (Cachia et al., 2026a,b), consistent with a broader literature that documents the efficacy of AI-delivered CBT chatbots for reducing depression and anxiety (Zhong, Luo and Zhang, 2024; Limpanopparat, Gibson and Harris, 2024; Heinz et al., 2025; Angelucci, Fabregas and Vazquez, 2026). We therefore expect a substantial first-stage effect (H1). This allows for credible tests of the downstream effects of mental health on intermediate psychological orientations (H2) and on liberal-democratic attitudes and out-group warmth (H3a, H3b).

The *Flourish* app is organized around a continuous support loop connecting goal-setting, real-time support, science-based activities, and reflection. Users begin by identifying personal well-being goals, such as managing anxiety, improving productivity, or deepening relationships. These goals provide context for subsequent one-tap mood check-ins and conversations with *Sunnie*, the AI well-being coach. Each interaction with *Sunnie* passes through a multi-stage process: a routing layer performs real-time risk detection and intent analysis, selects a specialized workflow (e.g., emotion regulation, goal-setting, venting, reflection), assembles longitudinal context from the user's goals, prior interactions, and past responses to different forms of support, and applies behavioral guardrails before a response reaches the user. Evidence-based techniques from CBT, DBT, ACT, positive psychology, motivational interviewing, and emotion-regulation research are embedded directly in these workflows and tools. *Sunnie* can recommend or guide activities, which users can also access through a science-based Toolkit, including brief journaling (Lieberman et al.,

[11] Open-ended responses will be coded using Anthropic's Claude Opus 5 and validated by two human coders.

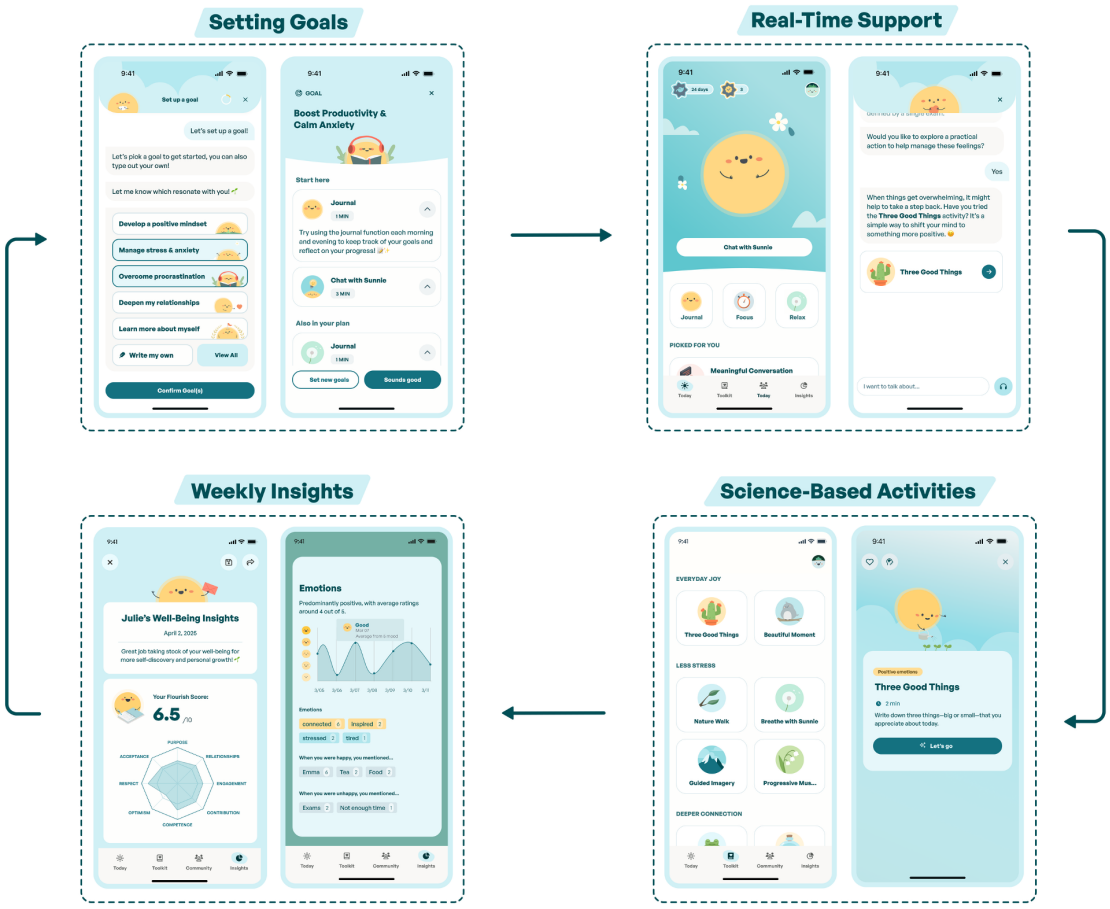


Figure 2: Continuous support loop of the *Flourish* app. Users set well-being goals, engage in check-ins and conversations with *Sunnie*, receive context-sensitive activity recommendations, and review personalized insights over time that help them reflect on progress and return to goal-setting with updated self-knowledge.

2007; Millgram et al., 2023; Torre and Lieberman, 2018), Three Good Things (Seligman et al., 2005), breathing exercises (Fincham et al., 2023; Hopper et al., 2019; Eskici İlgin and Yayla, 2023), progressive muscle relaxation (McCallie, Blum and Hood, 2006; Toussaint et al., 2021), and nature walks (Lee and Buchner, 2008; Ma, Lin and Williams, 2024; Nisbet, Zelenski and Grandpierre, 2019). Weekly insight reports summarize mood trends, emotional patterns, and progress, helping users reflect and return to goal-setting with updated self-knowledge. Figure 2 illustrates this loop with representative app screens.

For each treated participant, we observe aggregated usage metadata—specifically days active, hours active, number of chats with *Sunnie*, number of activities completed, number of journal

entries, and number of reward badges—linked through a unique participant identifier (the participant's *Flourish* Buddy ID; Section 3.3). We do not receive conversational content generated inside the application. This gives us granular compliance data without compromising participant privacy.

3.3 Incentive package and compliance

We invite respondents assigned to treatment, through the first engagement survey E1 (which doubles as the onboarding survey; SI Appendix A.4), to download *Flourish* and enter a fixed incentive package. The goal is to maximize compliance (app usage) over the six-week treatment window. The package consists of the following five components:

1. *Free premium access.* Three months of *Flourish*'s premium tier (normally $12.99 per month) at no cost.

2. *Paid engagement surveys.* Eight brief surveys (E1–E8) fielded across the six-week treatment window—two in week 1, one in each of weeks 2–5, and two in week 6. The first engagement survey (Day 8) is the onboarding survey: it asks the participant to consent, download *Flourish*, activate premium access with a study-specific code during the app's onboarding, and report their *Flourish* Buddy ID, which links their survey responses to their app usage data (SI Appendix A.4). E2–E8 each remind the respondent of the incentive structure, offer short guidance on effective app use, and deep-link into three in-app activities within *Flourish*, of which the respondent chooses one to complete; the set of three activities rotates across waves (SI Appendix A.5). Compensation for each engagement survey is contingent on completing one of the three linked in-app activities, which we verify through the usage metadata described above.[12] Each linked activity takes roughly 10 minutes to complete; completing all seven activity-linked surveys (E2–E8) thus implies approximately 70 minutes of in-app engagement. Deep links draw from the full set of in-app components (conversations with *Sunnie*, interactive journaling, and science-based Toolkit activities such as guided breathing and skill-practice modules).

[12]For this purpose, *Flourish* additionally provides a binary indicator of whether the account matched to a given Buddy ID has completed a given deep-linked activity.

3. *Flourished-days bonus.* A $20 bonus for reaching at least 20 days on the in-app "flourished days" counter during the six-week study period. The counter records the number of days on which participants complete at least one in-app activity.

4. *Weekly lottery.* Five $100 prizes per week for six weeks (30 winners, $3,000 total prize pool). Treated respondents accumulate lottery tickets in proportion to their number of flourished days. Tickets accrue weekly, but all six draws are conducted, and winners notified, only at disbursement, so no lottery outcome is revealed during the treatment or endline fielding windows.

5. *Reminders.* App-based push notifications plus email and text reminders from Forthright and *Flourish* throughout the treatment window.

We can directly observe compliance with the offer: each treated respondent's reported Buddy ID is matched to an app account with active premium access on the *Flourish* back-end (SI Appendix E). Conditional on completing onboarding, we expect high cumulative in-app usage given the engagement survey schedule described above. Our power analysis (Section 4.2) assumes that 65% of treated respondents complete onboarding.

We disburse all treatment-contingent payments—including lottery prizes, the flourished-days bonus, and engagement-survey compensation— as well as the control group placebo lottery payments, on a single fixed date after the endline fielding window closes, regardless of whether a participant completes the endline. We do not condition these payments on completing the endline, so the monetary incentive to complete it is the same in both arms. Both arms are paid for the screening, baseline, and endline surveys on completion, under Forthright's standard terms. We delay the treatment-contingent payments so that receiving money during the treatment window cannot itself move our outcomes. Recent evidence suggests that substantially larger cash transfers have only a modest direct effect on political attitudes (Broockman et al., 2024); delaying payment and introducing a placebo lottery remove this concern entirely.

3.4 Outcomes

We group our primary outcomes into four hypothesis families, ordered along the causal chain from mental health to political attitudes: mental health and well-being (H1), intermediate psychological orientations (H2), support for democratic norms (H3a), and attitudes toward out-groups (H3b).[13] Within each primary family, every subscale listed in Table 2 is a pre-registered primary outcome; most subscales aggregate multiple survey items into a single index. We document the full aggregation procedure below and in SI Appendix B. We use a Romano–Wolf step-down correction (Romano and Wolf, 2005) to hold the family-wise error rate at 0.05, applying it separately within H1, H2, H3a, and H3b. H1 gates the three downstream families (Section 4).

We construct each subscale in two steps. First, we reverse-code items as needed (matching the scoring convention of the source instrument) and z-standardize each item on the baseline distribution (pooled across arms, baseline wave only); this places items on a common scale before aggregation. Second, we take the equal-weighted mean of the item z-scores and re-z-standardize the subscale on the same baseline distribution, so that the subscale-level ITT is expressed in baseline-SD units. Table 2 provides an overview of the survey items and subscales. Full item wording and source citations are provided in the SI (Appendices A.2 and B).

Substantively, the three in-party norm-violation items in the H3a family capture core liberal-democratic principles—(i) respect for the rule of law (ignoring unfavorable court decisions), (ii) the separation of powers (circumventing Congress through executive orders), and (iii) press freedom (censoring critical media outlets)—but we deliberately operationalize each principle as a concrete violation that benefits the respondent's own party. Prior research shows that partisan motivations are central to how citizens trade off democratic principles—which most citizens endorse in the abstract—against concrete political goals: citizens tolerate, or fail even to recognize, violations committed on their party's behalf (Graham and Svolik, 2020; Simonovits, McCoy and Littvay, 2022; Krishnarajan, 2023). We note that this approach measures support for democratic norms as

[13] H2 collects four individual-level psychological orientations (zero-sum thinking, conversational receptiveness, political anxiety, and pessimism) that are neither mental-health batteries nor direct political attitudes; they are dispositions that plausibly respond to improved mental health and may in turn shape downstream political views.

they are contested in a polarized two-party system, rather than citizens' unconditional valuation of liberal-democratic principles (Ferrer et al., 2025).[14]

Table 2: Survey outcome measures by hypothesis family.

Subscale / battery	# Items	Sign
Mental health and well-being (H1)		
PHQ-8 depression	8	−
GAD-7 anxiety	7	−
UCLA loneliness	3	−
Perceived stress†	4	−
Negative affect (depression, anxiety, anger)	9	−
Flourishing	8	+
Positive affect (calm, well-being, vigor)	9	+
Intermediate psychological orientations (H2)		
Zero-sum thinking	4	−
Conversational receptiveness†	2	+
Political anxiety	4	−
Pessimism†	4	−
Support for democratic norms (H3a)		
In-party norm violations	3	−
Opposition to political violence†	1	+
Support for bipartisan cooperation	1	+
Attitudes toward out-groups (H3b)		
Generalized trust	1	+
Out-group thermometer (partisan-conditional)§	3	+
Affective polarization (in-party − out-party thermometer)‡	1	−
Secondary outcomes		
Political interest and turnout intention	2	
Out-party norm-compliance perception†	3	
Perceived out-party support for political violence	1	

Note: "Sign" gives the pre-registered direction of the treatment effect; secondary outcomes have no pre-registered direction. †Subscale or one or more constituent items reverse-coded so that higher subscale scores match the direction of the subscale's name. ‡In-party minus out-party feeling thermometer. §Partisan-conditional out-group thermometer; construction detailed in Section 4 and SI Appendix B.

[14]Generalized trust is not a group-specific measure of out-group attitudes. We retain it in H3b because the theoretical pathway from reduced distress through openness extends to generalized trust (Section 2.2), but note that it captures a related yet distinct dimension of out-group orientation: among partisans in the ANES data, the two thermometer-based H3b subscales are weakly correlated ($r = -0.28$), while generalized trust is not meaningfully correlated with either (SI Appendix C).

3.5 Ethics

We conduct the study under a protocol reviewed and approved by the Research Ethics Board at the first author's institution. All respondents provide active informed consent in the screening survey, fielded three days before baseline (SI Appendix A.1.1), based on a pre-approved script that covers the study purpose, voluntary participation, right to withdraw, compensation, data privacy practices, and risks of psychological discomfort from the mental-health batteries and the app. Participation is voluntary, and participants may withdraw at any wave without losing previously earned compensation. We remind participants, at every wave of engagement surveys, that *Flourish* is a wellness application, not a substitute for clinical care.

Four design choices further limit the study's ethical risks. First, to minimize psychological adversity for study participants, we employ an evidence-based technology that has been rigorously tested and shown to improve well-being without observed adverse effects (Cachia et al., 2026a,b). Second, we never see in-app content or identifiable information (Section 3.2): we receive only aggregated usage metadata, linked through the participant's *Flourish* Buddy ID, and never the conversations users have inside *Flourish*. Third, the two authors employed at *Flourish* are recused from implementation, data cleaning, analysis, and write-up, so the disclosed conflict of interest cannot shape the analysis. Finally, *Flourish* maintains a platform-level safety architecture: real-time risk detection, a constrained crisis-response workflow that surfaces region-appropriate crisis resources, and structured human review of de-identified medium- and high-risk interactions.

4 Empirical strategy

Our primary estimand is the ITT effect of being incentivized to use *Flourish* on each pre-specified subscale within the H1, H2, H3a, and H3b families. For every subscale, we estimate the specification in Equation 1 and report a one-sided p-value in the pre-registered direction. Inference within each family uses a Romano–Wolf step-down correction (Romano and Wolf, 2005) to control the family-wise error rate at 0.05, applied separately within H1, H2, H3a, and H3b. We test H2, H3a,

and H3b if the H1 estimates meet a first-stage threshold, i.e., if either of the following conditions holds: (i) for at least one H1 subscale, a separate RW-adjusted test rejects the null hypothesis that the ITT is at most 0.1 SD in the pre-registered direction (Section 4.1); or (ii) the cumulative ITT effect, summed in absolute value across those H1 subscales whose RW-adjusted tests reject a null of 0 in the pre-registered direction, exceeds 0.25 SD. If neither condition holds, the main text reports only the ITT and CACE estimates for the H1 outcomes and leaves out the analysis of all other outcomes. All tests use heteroskedasticity-robust (HC1) standard errors.

4.1 Inference

For respondent i in block b we estimate

$$Y_{ib}^{\text{post}} = \gamma_b + \tau Z_i + \lambda Y_{ib}^{\text{pre}} + \varepsilon_{ib}, \tag{1}$$

where Y_{ib}^{post} is the endline outcome, $Z_i \in \{0,1\}$ is treatment assignment, Y_{ib}^{pre} is the subscale-specific baseline outcome, and γ_b are block fixed effects for the 20 randomization blocks (Section 3.1). We estimate Equation 1 by OLS. We retain block fixed effects in all specifications because the randomization was stratified by block. The coefficient τ estimates the ITT; we test τ one-sided at $\alpha = 0.05$ in a pre-registered direction for every subscale (signs reported in Table 2 and explained in SI Appendix B). We expect baseline and endline scores to be highly correlated on mental-health and political batteries ($\rho \approx 0.8$; see Section 4.2), so controlling for Y_{ib}^{pre} buys considerable precision relative to a difference-in-means estimator. We will report a baseline-balance table comparing demographic characteristics and pre-treatment outcomes across treatment arms in the appendix, as well as a test for differential attrition employing the specification reported in Equation 1, where our outcome of interest takes the value of 1 if a baseline participant does not report any outcome of interest at endline. If we find a precisely estimated treatment effect on our attrition outcome ($p < 0.05$, two-tailed test), we will report sharp Lee-bounded estimates alongside our primary estimates.

We test each subscale one-sided in its pre-registered direction. Within H1, we predict a negative ITT on the five distress subscales (PHQ-8, GAD-7, UCLA loneliness, perceived stress, negative affect) and a positive ITT on the two well-being subscales (flourishing, positive affect). Within H2, we predict a negative ITT on zero-sum thinking, political anxiety, and pessimism, and a positive ITT on conversational receptiveness. Within H3a, we test each subscale in the direction consistent with greater support for liberal-democratic norms. Within H3b, we predict a positive ITT on generalized trust and the partisan-conditional out-group thermometer,[15] and a negative ITT on affective polarization. Subscale-specific signs are reported in the "Sign" column of Table 2 (see also SI Appendix B).

Identification assumptions. Identification of the ITT requires only random assignment and SUTVA. Random assignment is guaranteed by the within-block randomization described in Section 3.1. SUTVA requires no interference between units and a single treatment protocol. Spillovers from treated to control are highly unlikely because respondents generally are unknown to one another, and all treated respondents receive access to the same *Flourish* app. The additional assumptions required to interpret the CACE in Equation 2—a relevant first stage, monotonicity, and the exclusion restriction—are discussed in Section 4.3.

Adjusting for multiple outcomes. We correct for multiple testing within each primary family of outcomes using a Romano–Wolf step-down procedure (Romano and Wolf, 2005), which exploits the high intercorrelation among subscales and is less conservative than a Bonferroni correction. Within H1, we apply Romano–Wolf to the seven subscale p-values, controlling the family-wise error rate (FWER) at 0.05. We proceed to the three downstream families only if the H1 estimates meet the first-stage threshold stated at the beginning of this section; if so, we apply separate Romano–Wolf corrections to the four H2 subscale p-values, the three H3a subscale p-values, and

[15] All respondents rate eight thermometers at baseline and endline—Democrats, Republicans, LGBTQ people, undocumented immigrants in the US, Ivy-league academics, wealthy Americans (top 1%), gun owners, and evangelical Christians—and the pre-registered subscale averages the three out-group items for each respondent: LGBTQ people, undocumented immigrants, and Ivy-league academics for Republicans; wealthy Americans (top 1%), gun owners, and evangelical Christians for Democrats.

the three H3b subscale p-values, each at FWER 0.05. H1 gates the downstream families: without a meaningful first-stage effect on mental health, the H2 and H3 claims have no premise. For condition (i) of the threshold, we apply the same one-sided Romano–Wolf step-down to the seven H1 subscales as a separate family (FWER 0.05), with test statistics $(\hat{\tau}_s - 0.1)/\mathrm{SE}(\hat{\tau}_s)$, where $\hat{\tau}_s$ is the ITT estimate on subscale s from Equation 1. For this test, we reverse the sign of the estimates on the five distress subscales, so that positive values denote improvement. Once the threshold is met, we test H2, H3a, and H3b independently of one another. The downstream ITTs recover the total reduced-form effect of the *Flourish* offer on each subscale, not the distress-mediated pathway. We report all RW-adjusted p-values and point estimates in each family we test. We report secondary outcomes (political interest and turnout intention; out-party perceptions) with raw p-values; they are not part of the Romano–Wolf correction.

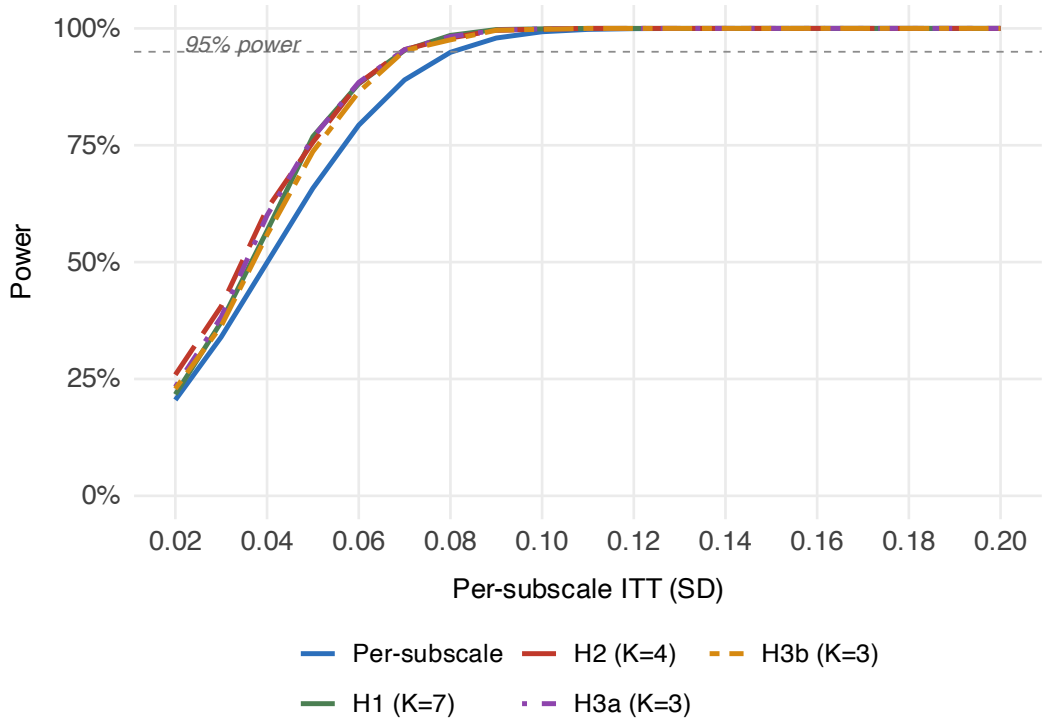
Attention checks. The study embeds three attention checks (SI Appendix A.3). The screening survey embeds one instructed-response row in the feeling-thermometer battery, asking respondents to move the slider to the top of the scale; respondents who do not answer exactly 100 are ineligible and are not invited to baseline. The baseline instrument embeds two further instructed-response items, in the Perceived stress and In-party norm violations batteries, each asking respondents to pick a specific option; we screen out respondents who fail either check before treatment assignment. The endline instrument has no attention check, so all attention-based exclusion occurs before treatment assignment.

Deviations from the pre-analysis plan. We will disclose any deviation from this pre-analysis plan in a subsection of the final paper titled "Deviations from the pre-analysis plan." For each one, we will report results under both the original pre-registered specification and the deviated one, so that readers can judge what the post-hoc choice did to the estimate.

4.2 Power analysis

We compute power by Monte Carlo simulation in the *DeclareDesign* framework (Blair et al., 2019). The simulation models the full data-generating process at $N_{\text{eff}} = 2{,}430$: baseline–endline correlation $\rho = 0.8$, 65% treated-arm compliance, 10% MCAR attrition, and the OLS specification in Equation 1 with HC1 standard errors, tested one-sided at $\alpha = 0.05$ in the pre-registered direction. For family-wise inference we extend the model to K correlated subscales and apply the Romano–Wolf step-down at within-family between-subscale correlation $\rho_{\text{sub}} = 0.5$. SI Appendix D gives the full data-generating process, the design-parameter table, and sensitivity analyses in ρ and ρ_{sub}.

Figure 3: Per-subscale and family-wise power as a function of the per-subscale ITT. All other parameters are fixed at the values in SI Appendix Table D.1.



Note: Power computed in *DeclareDesign* (Blair et al., 2019); see SI Appendix D for details. The solid blue curve shows per-subscale power at one-sided $\alpha = 0.05$ prior to correction. The family-wise curves show the probability that at least one subscale rejects under Romano–Wolf step-down in H1 ($K = 7$), H2 ($K = 4$), H3a ($K = 3$), and H3b ($K = 3$), with $\rho_{\text{sub}} = 0.5$; family-wise power exceeds per-subscale power because the family has K chances to reject. The horizontal dashed line marks 95% power.

Figure 3 plots power as a function of the true per-subscale ITT. For an ITT of 0.10 SD, per-

subscale power is 99.3% prior to correction and the per-subscale minimum detectable effect at 95% power is 0.080 SD. Under Romano–Wolf, family-wise power is at least 99% in each of the four hypothesis families (H1–H3b). Per-subscale power exceeds 95% so long as $\rho \geq 0.66$.

The first stage in our design—the ITT on mental health—is expected to be roughly 0.2–0.3 SD across H1 subscales, based on prior *Flourish* RCTs (Cachia et al., 2026a,b). The corresponding per-subscale MDE at 95% power is approximately 0.08 SD; we are therefore powered to detect downstream per-subscale effects on the order of one-third the first-stage ITT magnitude. Magnitudes in this range are well within those documented in similar RCTs. For example, Blattman, Jamison and Sheridan (2017) evaluate an 8-week CBT-informed program in Liberia and report ITT effects in the 0.2–0.3 SD range on both a positive-self-regard / mental-health index and a seven-item antisocial-behavior index, the primary downstream behavioral outcome. Our design is well powered to detect downstream effects of comparable magnitude.

4.3 Complier average causal effect

The ITT in Equation 1 estimates the effect of treatment assignment (being *offered Flourish*). To recover an effect of *using* the app among respondents induced to engage by the offer, we instrument a continuous measure of app engagement with random assignment and estimate a complier average causal effect (CACE). We define D_i as the total number of hours respondent i spent in *Flourish* over the six-week treatment window, drawn from the in-app usage logs (Section 3.2) (as a robustness check, we also report results that replace D_i with a binary onboarding indicator as the instrument-aligned compliance measure). Because control respondents receive no app access through the study, $D_i = 0$ by construction in the control arm, and the first-stage contrast is the treated arm's mean hours of use.[16] We obtain the CACE from

$$Y_{ib}^{\text{post}} = \gamma_b + \tau^{\text{CACE}} \hat{D}_i + \lambda Y_{ib}^{\text{pre}} + \varepsilon_{ib}, \tag{2}$$

[16] Usage data are linked through the *Flourish* Buddy IDs that treated respondents report in the onboarding survey; control respondents never receive that survey and are never asked for one, so measured usage is zero for controls by construction. *Flourish* is publicly available, and we assume that independent, off-study use by control respondents—who are never told the app's name during the treatment window—is negligible.

where $\hat{D}_i$ is the predicted value from the first-stage regression of D_i on Z_i, Y_{ib}^{pre}, and γ_b. Because Z_i is binary, 2SLS approximately reduces to the Wald ratio $\tau^{\text{CACE}} = \tau^{\text{ITT}}/E[D_i \mid Z_i = 1]$: the ITT on the outcome scaled by treated-arm mean hours of use. Under one-sided noncompliance and a linear treatment-response relationship between in-app hours and the outcome, τ^{CACE} is the per-hour effect on the outcome among respondents induced to engage by the offer.

Identification assumptions. Beyond random assignment and SUTVA (discussed in the identification assumptions component of Section 4), the CACE in Equation 2 requires three additional assumptions: a relevant first stage, monotonicity, and the exclusion restriction. The first stage is the mean hours of use for the treated arm, which we expect to be substantial given the incentive package described in Section 3.3. Monotonicity rules out defiers—respondents who would use the app if assigned to control but refuse it once encouraged to do so under treatment. We do not expect that being encouraged to use the app and offered free premium access would make any respondent less likely to use it than if they had to seek it out and pay for it themselves. The exclusion restriction requires that treatment assignment affects endline outcomes only through *Flourish* engagement. The two alternative channels we consider most plausible are engagement-survey compensation and lottery winnings received during the treatment window. Our design shuts both off: we disburse all treatment-contingent payments and reveal all lottery outcomes only after the endline fielding window closes (Section 3.3). Moreover, we provide lottery opportunities to respondents in both the treatment and control groups. This argument applies to the endline: because disbursement necessarily precedes the conditional follow-up wave (Section 3), we do not extend the exclusion restriction to CACE estimates on the follow-up. We nonetheless expect any direct effect of compensation on follow-up outcomes to be negligible: the typical treated respondent receives $32 in cash—$12 across the eight engagement surveys and the $20 flourished-days bonus—plus three months of free app access (a $39 value), and endowments of this magnitude—or substantially larger—have not been found to shift well-being (Jha and Shayo, 2019) or political attitudes (Broockman et al., 2024).

4.4 Heterogeneity and secondary analyses

Pre-specified CATEs. We pre-register three subgroup conditional average treatment effects (CATEs) for every primary subscale in the H1, H2, H3a, and H3b families: by (i) gender (male/female), (ii) partisanship (Democrat/Republican), and (iii) baseline mental health (above vs. below median of the H1 baseline mental-health index used for randomization blocking; Section 3.1).[17] We estimate each CATE by fitting Equation 1 on the relevant subsample (e.g., men only, Democrats only, above-median baseline mental health only) rather than by adding treatment-by-subgroup interactions to the full-sample specification. We pre-register these CATEs for transparency. They are not part of the confirmatory H1–H3b tests, and their p-values do not enter any Romano–Wolf correction.

Dosage heterogeneity. Within the treated arm, we examine whether heavier in-app use predicts better endline outcomes. Specifically, we regress each primary subscale on total in-app hours accumulated over the six-week window (from the *Flourish* usage logs; Section 3.2), controlling for the baseline outcome Y_{ib}^{pre} and block fixed effects γ_b, and report the slope per hour of use. Hours of use is post-treatment, so this estimate is not causal: block fixed effects do not absorb selection into engagement, and heavy users may differ from light users on unobserved traits that also predict the outcome. The estimate describes how observed engagement covaries with endline outcomes. It is not a test of H1, H2, H3a, or H3b, and we report it as exploratory alongside the pre-specified CATEs.

Exploratory mediation analysis. We estimate causal mediation models for the subscales on which we detect treatment effects, linking intermediate psychological orientations (H2) to political outcomes in the H3a and H3b families. For each mediator–outcome pair, taken one at a time, we estimate the average causal mediation effect (ACME) and the average direct effect in the framework of Imai et al. (2011), conditioning on block fixed effects and the baseline values of the mediator and the outcome. Because the mediators are not randomly assigned, the sequential-ignorability

[17] Nonbinary respondents and those who self-describe are not included in this CATE.

assumption underlying these estimates is unlikely to hold in our design (Green, Ha and Bullock, 2010). We therefore treat any results from this analysis as suggestive evidence on mechanisms and report them in the SI Appendix.

Secondary outcomes. We pre-specify three exploratory secondary outcomes that we estimate using Equation 1 and report with unadjusted two-sided p-values: political interest and turnout intention, and two subscales capturing perceptions of the out-party—perceived out-party compliance with democratic norms and perceived out-party support for political violence. The theoretical case for downstream effects on out-party perceptions is more ambiguous than for the affective H3b measures; to the extent that the intervention reduces anxiety and threat perceptions, however, treated respondents may become less likely to assume the worst about out-partisans. Secondary outcomes are not pre-registered with a directional hypothesis and are not part of any of the H1–H3b testing procedures.

Descriptive analyses of usage data. Using *Flourish* log data, we will report descriptive statistics on overall usage of the app (i.e., time spent on the app), as well as usage of specific app features.

5 Results

To be completed following implementation of the pre-registered analysis plan (Section 4).

6 Discussion

To be completed following implementation.

7 Tentative conclusion

In this registered report, we consider the causal effects of mental health on support for the underlying foundations of liberal democracy, and evaluate the potential of mental health interventions to increase support for democracy. To do so, we develop a theoretical framework that clarifies why improvements to mental health may shape political dispositions, which, in turn, shape democratic attitudes. To test the observable implications of our theoretical framework, we designed a large-scale RCT in collaboration with *Flourish*, in which we encourage study participants to use the *Flourish* well-being application for six weeks, and elicit social and political outcomes in seemingly unrelated surveys.

This registered report is notable for three reasons. First, it provides design-based causal evidence on the effects of psychological distress on support for the underlying foundations of liberal democracy. This evidence complements existing research documenting associations between distress and adverse political outcomes (Institute for Policy Research, 2023; Baum et al., 2025; Ojeda, 2025; Peterson et al., 2025; Ward et al., 2025), while clarifying the extent to which published correlational evidence on this topic should be interpreted causally. Second, we take a broad measurement approach, examining how different forms of distress shape multiple dimensions of support for liberal democracy. In doing so, we complement studies that focus on specific empirical relationships—such as depression and belief in conspiracy theories (Green et al., 2023)—and provide a rich perspective on the complex relationship between mental health and democratic citizenship. Finally, our RCT speaks not only to theoretical debates in the field, but also to pressing policy challenges. By collaborating with *Flourish* and evaluating the potential positive externalities of an increasingly popular technology, we examine whether mental health interventions can improve personal well-being and, in doing so, generate broader societal consequences by strengthening support for the foundations of liberal democracy.

Our study contributes to three lines of research. First, we contribute to work on the psychological foundations of support for liberal democracy (Baum et al., 2024; Albertson and Gadarian, 2015;

Baum et al., 2025; Balmas, Atia and Halperin, 2023; Green et al., 2023). Specifically, we develop a comprehensive theoretical framework to explain why improved mental health may bolster support for democracy, and test its observable implications through a novel RCT, providing the first causal evidence on the relationship between mental health and support for democracy. Second, we contribute to research on interventions designed to reduce polarization and strengthen support for democracy (Voelkel et al., 2024; Weiss, Green and Willer, Forthcoming; Hartman et al., 2022; Druckman et al., 2023) by evaluating a non-paternalistic, incentive-compatible approach that aims to improve personal well-being, while generating positive democratic externalities. This approach differs from many light-touch interventions, whose effects may be small, fleeting, or limited by citizens' unwillingness to engage with counter-attitudinal content (Holliday, Lelkes and Westwood, 2025; Landry and Halperin, 2025; Weiss, Siegel and Scacco, Forthcoming). Finally, we contribute to research on the political consequences of emerging technologies (Magistro et al., 2026; Borwein et al., 2026; Zilinsky and Zeitzoff, 2024), by examining whether AI-powered mental health support—a scalable technology that may complement human care, especially for underserved populations—has downstream consequences for democratic health.

References

Albertson, Bethany and Shana Kushner Gadarian. 2015. *Anxious Politics: Democratic Citizenship in a Threatening World*. New York: Cambridge University Press.

American National Election Studies. 2026. "ANES 2024 Time Series Study Full Release." https://electionstudies.org/data-center/2024-time-series-study/. May 19, 2026 version [dataset and documentation].

Andersson, Gerhard, Per Carlbring, Nickolai Titov and Nils Lindefors. 2019. "Internet interventions for adults with anxiety and mood disorders: a narrative umbrella review of recent meta-analyses." *The Canadian Journal of Psychiatry* 64(7):465–470.

Andrews Fearon, Patricia and Friedrich M. Götz. 2024. "The zero-sum mindset." *Journal of Personality and Social Psychology* 127(4):758–795.

Angelucci, Manuela, Raissa Fabregas and Antonia Vazquez. 2026. The Well-Being Effects of Digital Mental Health Care. IZA Discussion Paper 18538 Institute of Labor Economics (IZA) Bonn: .

URL: *https://www.iza.org/publications/dp/18538/the-well-being-effects-of-digital-mental-health-care*

Arceneaux, Kevin. 2017. "Anxiety reduces empathy toward outgroup members but not ingroup members." *Journal of Experimental Political Science* 4(1):68–80.

Armingeon, Klaus and Kai Guthmann. 2014. "Democracy in crisis? The declining support for national democracy in E uropean countries, 2007–2011." *European Journal of Political Research* 53(3):423–442.

Asimovic, Nejla, Jonathan Nagler, Richard Bonneau and Joshua A Tucker. 2021. "Testing the effects of Facebook usage in an ethnically polarized setting." *Proceedings of the National Academy of Sciences* 118(25):e2022819118.

Baarck, Julia, Béatrice d'Hombres and Guido Tintori. 2022. "Loneliness in Europe before and during the COVID-19 pandemic." *Health policy* 126(11):1124–1129.

Balmas, Meital, Renana Atia and Eran Halperin. 2023. "The Mediating Role of Depression in the Relationship Between News Consumption and Interparty Hostility During Covid-19." *International Journal of Communication* 17:304–330.

Balmas, Meital, Tal Orian Harel and Eran Halperin. 2022. "I hate you when I am anxious: Anxiety during the COVID-19 epidemic and ideological hostility." *Journal of Applied Social Psychology* 52(11):1081–1093.

Bar-Haim, Yair, Dominique Lamy, Lee Pergamin, Marian J Bakermans-Kranenburg and Marinus H Van Ijzendoorn. 2007. "Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study." *Psychological bulletin* 133(1):1.

Baum, Matthew A, James N Druckman, Katherine Ognyanova, David Lazer and Roy H Perlis. 2025. "Depression, Populism, and Presidential Approval." *International Journal of Public Opinion Research* 37(4):edaf050.

Baum, Matthew A, James N Druckman, Matthew D Simonson, Jennifer Lin and Roy H Perlis. 2024. "The political consequences of depression: how conspiracy beliefs, participatory inclinations, and depression affect support for political violence." *American journal of political science* 68(2):575–594.

Beck, Aaron T. 1970. "Cognitive therapy: Nature and relation to behavior therapy." *Behavior therapy* 1(2):184–200.

Beck, Aaron T and Brad A Alford. 2014. Depression: Causes and treatment. In *Depression*. University of Pennsylvania Press.

Bell, Victoria, Benjamin Robinson, Cornelius Katona, Anne-Kathrin Fett and Sukhi Shergill. 2019.

"When trust is lost: The impact of interpersonal trauma on social interactions." *Psychological medicine* 49(6):1041–1046.

Berlingieri, Francesco, Béatrice d'Hombres and Matija Kovacic. 2025. Disentangling loneliness and trust in populist voting behaviour in Europe. Technical report GLO Discussion Paper.

Blair, Graeme, Alexander Coppock and Margaret Moor. 2020. "When to worry about sensitivity bias: A social reference theory and evidence from 30 years of list experiments." *American Political Science Review* 114(4):1297–1315.

Blair, Graeme, Jasper Cooper, Alexander Coppock and Macartan Humphreys. 2019. "Declaring and diagnosing research designs." *American Political Science Review* 113(3):838–859.

Blattman, Christopher, Julian C. Jamison and Margaret Sheridan. 2017. "Reducing crime and violence: Experimental evidence from cognitive behavioral therapy in Liberia." *American Economic Review* 107(4):1165–1206.
URL: *https://doi.org/10.1257/aer.20150503*

Blattman, Christopher, Sebastian Chaskel, Julian C. Jamison and Margaret Sheridan. 2023. "Cognitive Behavioral Therapy Reduces Crime and Violence over Ten Years: Experimental Evidence." *American Economic Review: Insights* 5(4):527–545.
URL: *https://doi.org/10.1257/aeri.20220427*

Bolier, Linda, Merel Haverman, Gerben J Westerhof, Heleen Riper, Filip Smit and Ernst Bohlmeijer. 2013. "Positive psychology interventions: A meta-analysis of randomized controlled studies." *BMC public health* 13(1):119.

Borawski, Dominik. 2018. "The loneliness of the zero-sum game loser. The balance of social exchange and belief in a zero-sum game as predictors of loneliness." *Personality and Individual Differences* 135:270–276.

Borwein, Sophie, Beatrice Magistro, R Michael Alvarez, Bart Bonikowski and Peter J Loewen. 2026. "Explaining womens skepticism toward artificial intelligence: The role of risk orientation and risk exposure." *PNAS nexus* 5(1):pgaf399.

Britton, Jennifer C, Shmuel Lissek, Christian Grillon, Maxine A Norcross and Daniel S Pine. 2011. "Development of anxiety: The role of threat appraisal and fear learning." *Depression and anxiety* 28(1):5–17.

Broockman, David E, Elizabeth Rhodes, Alexander W Bartik, Karina Dotson, Sarah Miller, Patrick K Krause and Eva Vivalt. 2024. The causal effects of income on political attitudes and behavior: A randomized field experiment. Technical report National Bureau of Economic Research.

Broockman, David E and Joshua L Kalla. 2025. "Consuming cross-cutting media causes learning and moderates attitudes: A field experiment with Fox News viewers." *The Journal of Politics* 87(1):246–261.

Broockman, David E, Joshua L Kalla and Sean J Westwood. 2023. "Does affective polarization undermine democratic norms or accountability? Maybe not." *American Journal of Political Science* 67(3):808–828.

Cachia, Julie Y. A., Xuan Zhao, John Hunter, Delancey Wu, Eta Lin and Julian De Freitas. 2026a. "AI for Proactive Mental Health: A Multi-Institutional, Longitudinal Randomized Controlled Trial." *NEJM AI* 3(8).
URL: *https://ai.nejm.org/doi/full/10.1056/AIoa2501293*

Cachia, Julie Y. A., Xuan Zhao, Wenxin Pu, Chen Liao, Merel Keijsers, Julian De Freitas and Rebecca Fortgang. 2026b. "AI-Powered Mental Health Intervention for Depression and Anxiety: A Randomized Clinical Trial." PsyArXiv preprint.
URL: *https://osf.io/preprints/psyarxiv/nwz9q_v5*

Chen, Xing and Xiaoxiao Shen. 2025. "Silent suffering: Depression, perceived inequality, and political trust." *Research & Politics* 12(3):20531680251374167.

Chinoy, Sahil, Nathan Nunn, Sandra Sequeira and Stefanie Stantcheva. 2026. "Zero-Sum Thinking and the Roots of US Political Differences." *American Economic Review* 116(3):1052–1096.

Claassen, Christopher and Pedro C Magalhães. 2023. "Public support for democracy in the United States has declined generationally." *Public opinion quarterly* 87(3):719–732.

Clement, Sarah, Oliver Schauman, Tanya Graham, Francesca Maggioni, Sara Evans-Lacko, Nikita Bezborodovs, Craig Morgan, Nicolas Rüsch, June SL Brown and Graham Thornicroft. 2015. "What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies." *Psychological medicine* 45(1):11–27.

Cohen, Sheldon. 1988. "Perceived stress in a probability sample of the United States.".

Cohen, Sheldon, Tom Kamarck and Robin Mermelstein. 1983. "A global measure of perceived stress." *Journal of Health and Social Behavior* 24(4):385–396.

Daly, Michael and Lucía Macchia. 2023. "Global trends in emotional distress." *Proceedings of the National Academy of Sciences* 120(14):e2216207120.

Diener, Ed, Derrick Wirtz, William Tov, Chu Kim-Prieto, Dong-won Choi, Shigehiro Oishi and Robert Biswas-Diener. 2010. "New well-being measures: Short scales to assess flourishing and positive and negative feelings." *Social Indicators Research* 97(2):143–156.

Druckman, James N. 2024. "How to study democratic backsliding." *Political Psychology* 45:3–42.

Druckman, James N, Suji Kang, James Chu, Michael N. Stagnaro, Jan G Voelkel, Joseph S Mernyk, Sophia L Pink, Chrystal Redekopp, David G Rand and Robb Willer. 2023. "Correcting misperceptions of out-partisans decreases American legislators support for undemocratic practices." *Proceedings of the National Academy of Sciences* 120(23):e2301836120.

Dubois, Hans. 2025. Mental health: Risk groups, trends, services and policies. Research Report EF25029 European Foundation for the Improvement of Living and Working Conditions Luxembourg: .

URL: *https://eurofound.europa.eu/en/publications/all/mental-health-risk-groups-trends-services-and-policies*

Edinger, Andy, Johan Bollen, Hernán A Makse and Matteo Serafino. 2025. "Cognitive distortions are associated with increasing political polarization." *Communications Psychology* 3(1):105.

Eskici İlgin, Vesile and Ayşegül Yayla. 2023. "Effect of the 4-7-8 breathing technique on pain level and sleep quality of patients after laparoscopic bariatric surgery: A randomized controlled study." *Bariatric Surgical Practice and Patient Care* 18(4):225–232.

Ferrer, Sergi, Enrique Hernández, Enrique Prada and Damjan Tomic. 2025. "The value of liberal democracy: Assessing citizens' commitment to democratic principles." *European journal of political research* 64(4):2018–2038.

Fincham, Guy William, Clara Strauss, Jesus Montero-Marin and Kate Cavanagh. 2023. "Effect of breathwork on stress and mental health: A meta-analysis of randomised-controlled trials." *Scientific reports* 13(1):432.

Franco, Annie, Neil Malhotra and Gabor Simonovits. 2014. "Publication bias in the social sciences: Unlocking the file drawer." *Science* 345(6203):1502–1505.

Fréget, Louis, Ariell Reshef and Claudia Senik. 2026. "Do Personal AI Conversations Reduce Loneliness? Evidence from a Pre-Registered Experiment.".

García-Ponce, Omar, Lauren E Young and Thomas Zeitzoff. 2023. "Anger and support for retribution in Mexicos drug war." *Journal of Peace Research* 60(2):274–290.

Graham, Carol. 2024. "Hope, Despair, and Long-Term Outcomes.". Accessed: 2026-04-24.

URL: *https://www.brookings.edu/articles/hope-despair-and-long-term-outcomes/*

Graham, Matthew H and Milan W Svolik. 2020. "Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States." *American Political Science Review* 114(2):392–409.

Green, Donald P, Shang E Ha and John G Bullock. 2010. "Enough already about black box experiments: Studying mediation is more difficult than most scholars suppose." *The Annals of the American Academy of Political and Social Science* 628(1):200–208.

Green, Jon, James N Druckman, Matthew A Baum, David Lazer, Katherine Ognyanova and Roy H Perlis. 2023. "Depressive symptoms and conspiracy beliefs." *Applied Cognitive Psychology* 37(2):332–359.

Hanlon, Michelle, Saumitra Jha, Namrata Kala, Nemit Shroff and Chagai M Weiss. 2026. "Seeing Green: The Effects of Financial Exposures on Support for Climate Action." *NBER* .

Hartman, Rachel, Will Blakey, Jake Womick, Chris Bail, Eli J Finkel, Hahrie Han, John Sarrouf, Juliana Schroeder, Paschal Sheeran, Jay J Van Bavel et al. 2022. "Interventions to reduce partisan animosity." *Nature human behaviour* 6(9):1194–1205.

Haushofer, Johannes, Sara Lowes, Abednego Musau, David Ndetei, Nathan Nunn, Moritz Poll and Nancy Qian. 2023. "Stress, ethnicity, and prosocial behavior." *Journal of Political Economy Microeconomics* 1(2):225–269.

Hayes, Steven C, Jason B Luoma, Frank W Bond, Akihiko Masuda and Jason Lillis. 2006. "Acceptance and commitment therapy: Model, processes and outcomes." *Behaviour research and therapy* 44(1):1–25.

Heinz, Michael V., Daniel M. Mackin, Brianna M. Trudeau, Sukanya Bhattacharya, Yinzhuo Wang, Haley A. Banta, Abi D. Jewett, Abigail J. Salzhauser, Tess Z. Griffin and Nicholas C. Jacobson. 2025. "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment." *NEJM AI* 2(4):AIoa2400802.

URL: *https://ai.nejm.org/doi/10.1056/AIoa2400802*

Heller, Sara B, Anuj K Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan and Harold A Pollack. 2017. "Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago." *The Quarterly Journal of Economics* 132(1):1–54.

Hirsch-Hoefler, Sivan, Daphna Canetti, Carmit Rapaport and Stevan E. Hobfoll. 2016. "Conflict Will Harden Your Heart: Exposure to Violence, Psychological Distress, and Peace Barriers in Israel and Palestine." *British Journal of Political Science* 46(4):845–859.

Holliday, Derek E, Shanto Iyengar, Yphtach Lelkes and Sean J Westwood. 2024. "Uncommon and nonpartisan: Antidemocratic attitudes in the American public." *Proceedings of the National Academy of Sciences* 121(13):e2313013121.

Holliday, Derek E, Yphtach Lelkes and Sean J Westwood. 2025. "Why depolarization is hard: Evaluating attempts to decrease partisan animosity in America." *Proceedings of the National Academy of Sciences* 122(39):e2508827122.

Hopper, Susan I, Sherrie L Murray, Lucille R Ferrara and Joanne K Singleton. 2019. "Effectiveness of diaphragmatic breathing for reducing physiological and psychological stress in adults: a quantitative systematic review." *JBI Evidence Synthesis* 17(9):1855–1876.

Huddy, Leonie, Stanley Feldman and Christopher Weber. 2007. "The political consequences of perceived threat and felt insecurity." *The ANNALS of the American Academy of Political and Social Science* 614(1):131–153.

Hughes, Mary Elizabeth, Linda J Waite, Louise C Hawkley and John T Cacioppo. 2004. "A short scale for measuring loneliness in large surveys: Results from two population-based studies." *Research on aging* 26(6):655–672.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4):765–789.

Institute for Policy Research, Northwestern University. 2023. "The Political Consequences of Poor Mental Health.". Accessed: 2026-04-24.
URL: *https://www.ipr.northwestern.edu/news/2023/the-political-consequences-of-poor-mental-health.html*

Iyengar, Shanto and Sean J Westwood. 2015. "Fear and loathing across party lines: New evidence on group polarization." *American journal of political science* 59(3):690–707.

Jenkins, Brooke N, Marie P Cross, Candice D Donaldson, Sarah D Pressman, Michelle A Fortier, Zeev N Kain, Sheldon Cohen, Logan T Martin and George Farkas. 2023. "The subcomponents of affect scale (SAS): validating a widely used affect scale." *Psychology & health* 38(8):1032–1055.

Jha, Saumitra and Moses Shayo. 2019. "Valuing peace: the effects of financial market exposure on votes and political attitudes." *Econometrica* 87(5):1561–1588.

Jha, Saumitra, Moses Shayo and Chagai M Weiss. 2025. "Financial market exposure increases generalized trust." *Journal of Public Economics* 242:105303.

Kannan, Viji Diane and Peter J Veazie. 2023. "US trends in social isolation, social engagement, and companionship—nationally and by age, sex, race/ethnicity, family income, and work hours, 2003–2020." *SSM-population health* 21:101331.

Kazdin, Alan E and Stacey L Blase. 2011. "Rebooting psychotherapy research and practice to reduce the burden of mental illness." *Perspectives on psychological science* 6(1):21–37.

Kliem, Sören, Christoph Kröger and Joachim Kosfelder. 2010. "Dialectical behavior therapy for borderline personality disorder: a meta-analysis using mixed-effects modeling." *Journal of consulting and clinical psychology* 78(6):936.

Krishnarajan, Suthan. 2023. "Rationalizing democracy: The perceptual bias and (un)democratic behavior." *American Political Science Review* 117(2):474–496.

Kroenke, Kurt, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry and Ali H. Mokdad. 2009. "The PHQ-8 as a Measure of Current Depression in the General Population." *Journal of Affective Disorders* 114(1–3):163–173.

Ladd, Jonathan McDonald and Gabriel S Lenz. 2008. "Reassessing the role of anxiety in vote choice." *Political Psychology* 29(2):275–296.

Landry, Alexander P and Eran Halperin. 2025. "Intergroup psychological interventions: The motivational challenge." *American Psychologist* 80(2):206.

Landwehr, Claudia and Christopher Ojeda. 2021. "Democracy and depression: A cross-national study of depressive symptoms and nonparticipation." *American Political Science Review* 115(1):323–330.

Lee, Amber Hye-Yon. 2022. "Social trust in polarized times: How perceptions of political polarization affect Americans trust in each other." *Political behavior* 44(3):1533–1554.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies* 76(3):1071–1102.
URL: *http://www.jstor.org/stable/40247633*

Lee, I-Min and David M Buchner. 2008. "The importance of walking to public health." *Medicine & science in sports & exercise* 40(7):S512–S518.

Levitsky, Steven and Daniel Ziblatt. 2019. *How democracies die*. Crown.

Lieberman, Matthew D, Naomi I Eisenberger, Molly J Crockett, Sabrina M Tom, Jennifer H Pfeifer and Baldwin M Way. 2007. "Putting feelings into words." *Psychological science* 18(5):421–428.

Limpanopparat, Sucharat, Erin Gibson and Andrew Harris. 2024. "User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: A systematic review." *Computers in Human Behavior: Artificial Humans* 2(2):100081.

Linardon, Jake, Pim Cuijpers, Per Carlbring, Mariel Messer and Matthew Fuller-Tyszkiewicz. 2019. "The efficacy of app-supported smartphone interventions for mental health problems: A meta-analysis of randomized controlled trials." *World Psychiatry* 18(3):325–336.

Linehan, Marsha M and Chelsey R Wilks. 2015. "The course and evolution of dialectical behavior therapy." *American journal of psychotherapy* 69(2):97–110.

Liu, Zhuo and Mengni Kuai. 2025. "The global burden of depression in adolescents and young adults, 1990–2021: systematic analysis of the global burden of disease study." *BMC psychiatry* 25(1):767.

Ma, Jingni, Peilin Lin and Joanne Williams. 2024. "Effectiveness of nature-based walking interventions in improving mental health in adults: a systematic review." *Current Psychology* 43(11):9521–9539.

Magistro, Beatrice, Sophie Borwein, R Michael Alvarez, Bart Bonikowski and Peter John Loewen. 2026. "Attitudes toward artificial intelligence (AI) and globalization: Common microfoundations and political implications." *American Journal of Political Science* 70(1):348–365.

McCallie, Martha S, Claire M Blum and Charlaine J Hood. 2006. "Progressive muscle relaxation." *Journal of human behavior in the social environment* 13(3):51–66.

McClendon, Gwyneth H. 2014. "Social esteem and participation in contentious politics: A field experiment at an LGBT pride rally." *American Journal of Political Science* 58(2):279–290.

McGrath, John J, Ali Al-Hamzawi, Jordi Alonso, Yasmin Altwaijri, Laura H Andrade, Evelyn J Bromet, Ronny Bruffaerts, José Miguel Caldas de Almeida, Stephanie Chardoul, Wai Tat Chiu et al. 2023. "Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries." *The Lancet Psychiatry* 10(9):668–681.

Millgram, Yael, Matthew K Nock, David D Bailey and Amit Goldenberg. 2023. "Knowledge about

the source of emotion predicts emotion-regulation attempts, strategies, and perceived emotion-regulation success." *Psychological Science* 34(11):1244–1255.

Minson, Julia A. and Frances S. Chen. 2022. "Receptiveness to Opposing Views: Conceptualization and Integrative Review." *Personality and Social Psychology Review* 26(2):93–111.

Moniz, Philip, James N Druckman and Jeremy Freese. 2025. "The file drawer problem in social science survey experiments." *Proceedings of the National Academy of Sciences* 122(12):e2426937122.

Mudde, Cas. 2007. *Populist radical right parties in Europe*. Vol. 22 Cambridge university press Cambridge.

Nisbet, Elizabeth K, John M Zelenski and Zsuzsa Grandpierre. 2019. "Mindfulness in nature enhances connectedness and mood." *Ecopsychology* 11(2):81–91.

Office of the Surgeon General. 2023. Our epidemic of loneliness and isolation: The US Surgeon Generals Advisory on the healing effects of social connection and community. Technical report US Department of Health and Human Services.

Ojeda, Christopher. 2025. *The sad citizen: How politics is depressing and why it matters*. University of Chicago Press.

Ojeda, Christopher, Luca Bernardi and Claudia Landwehr. 2023. "Depression and the gender gap in political interest." *Electoral Studies* 82:102598.

Peterson, Delaney, Matthijs Rooduijn, Frederic R. Hopp, Gijs Schumacher and Bert N. Bakker. 2025. "Loneliness is positively associated with populist radical right support." *Social Science & Medicine* 366:117676.
URL: *https://doi.org/10.1016/j.socscimed.2025.117676*

Pew Research Center. 2017. Americans Express Increasingly Warm Feelings Toward Religious Groups. Technical report Pew Research Center.

URL: *https://www.pewresearch.org/wp-content/uploads/sites/20/2017/02/Feeling-thermometer-report-FOR-WEB.pdf*

Piao, Xiangdan, Jun Xie and Shunsuke Managi. 2024. "Continuous worsening of population emotional stress globally: universality and variations." *BMC Public Health* 24(1):3576.

Polarization Research Lab. 2024. "America's Political Pulse." https://americaspoliticalpulse.com/citizens/. Dartmouth College, University of Pennsylvania, and Stanford University.

Romano, Joseph P. and Michael Wolf. 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association* 100(469):94–108.

Różycka-Tran, Joanna, Jarosław P Piotrowski, Magdalena Żemojtel-Piotrowska, Paweł Jurek, Evgeny N Osin, Byron G Adams, Rahkman Ardi, Sergiu Bălătescu, Arbinda Lal Bhomi, Sergey A Bogomaz et al. 2021. "Belief in a zero-sum game and subjective well-being across 35 countries." *Current Psychology* 40(7):3575–3584.

Scheier, Michael F, Charles S Carver and Michael W Bridges. 1994. "Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test." *Journal of personality and social psychology* 67(6):1063.

Seligman, Martin EP, Tracy A Steen, Nansook Park and Christopher Peterson. 2005. "Positive psychology progress: empirical validation of interventions." *American psychologist* 60(5):410.

Shri, Richa. 2010. "Anxiety: causes and management." *The Journal of Behavioral Science* 5(1):100–118.

Simonovits, Gábor, Jennifer McCoy and Levente Littvay. 2022. "Democratic hypocrisy and out-group threat: Explaining citizen support for democratic erosion." *The Journal of Politics* 84(3):1806–1811.

Sin, Nancy L and Sonja Lyubomirsky. 2009. "Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: A practice-friendly meta-analysis." *Journal of clinical psychology* 65(5):467–487.

Smith, Kevin B., Aaron C. Weinschenk and Costas Panagopoulos. 2024. "On pins and needles: anxiety, politics and the 2020 U.S. Presidential election." *Journal of Elections, Public Opinion and Parties* 34(3):409–426.

Spitzer, Robert L., Kurt Kroenke, Janet B. W. Williams and Bernd Löwe. 2006. "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7." *Archives of Internal Medicine* 166(10):1092–1097.

Steenvoorden, Eefje and Eelco Harteveld. 2018. "The appeal of nostalgia: The influence of societal pessimism on support for populist radical right parties." *West European Politics* 41(1):28–52.

Stefaniak, A., R. K. Mallett and M. J. A. Wohl. 2020. "Zero-sum beliefs shape advantaged allies' support for collective action." *European Journal of Social Psychology* 50(6):1259–1275.

Stickley, Andrew, Tomiki Sumiyoshi, Naoki Kondo, Mall Leinsalu, Yosuke Inoue, Vladislav Ruchkin, Jae Il Shin and Martin McKee. 2023. "Psychological distress and voting behaviour in nine countries of the former Soviet Union." *Scientific reports* 13(1):22709.

Sullivan, John L, James Piereson and George E Marcus. 1993. *Political tolerance and American democracy*. University of Chicago Press.

Thompson, Derek. 2025. "The Anti-Social Century." *The Atlantic* . Accessed: 2026-04-24.
URL: *https://www.theatlantic.com/magazine/archive/2025/02/american-loneliness-personality-politics/681091/*

Torre, Jared B and Matthew D Lieberman. 2018. "Putting feelings into words: Affect labeling as implicit emotion regulation." *Emotion Review* 10(2):116–124.

Toussaint, Loren, Quang Anh Nguyen, Claire Roettger, Kiara Dixon, Martin Offenbächer, Niko Kohls, Jameson Hirsch and Fuschia Sirois. 2021. "Effectiveness of progressive muscle relaxation, deep breathing, and guided imagery in promoting psychological and physiological states of relaxation." *Evidence-Based Complementary and Alternative Medicine* 2021(1):5924040.

Valentim, Vicente. 2024. *The normalization of the radical right: A norms theory of political supply and demand*. Oxford University Press.

Valentim, Vicente, Elias Dinas and Daniel Ziblatt. 2025. "How mainstream politicians Erode norms." *British Journal of Political Science* 55:e105.

Valgarðsson, Viktor, Will Jennings, Gerry Stoker, Hannah Bunting, Daniel Devine, Lawrence McKay and Andrew Klassen. 2025. "A crisis of political trust? Global trends in institutional trust from 1958 to 2019." *British Journal of Political Science* 55:e15.

Vasilopoulos, Pavlos, George E Marcus and Martial Foucault. 2018. "Emotional responses to the Charlie Hebdo attacks: Addressing the authoritarianism puzzle." *Political psychology* 39(3):557–575.

Voelkel, Jan G, Michael N Stagnaro, James Y Chu, Sophia L Pink, Joseph S Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi G Allen et al. 2024. "Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity." *Science* 386(6719):eadh4764.

Vries, Catherine E De and Sara Hobolt. 2020. *Political entrepreneurs: The rise of challenger parties in Europe*. Princeton University Press.

Wagner, Markus and Davide Morisi. 2019. Anxiety, fear, and political decision making. In *Oxford research encyclopedia of politics*. Oxford University Press.

Ward, George, H Andrew Schwartz, Salvatore Giorgi, Jochen I Menges and Sandra C Matz. 2025.

"The role of negative affect in shaping populist support: Converging field evidence from across the globe." *American Psychologist* 80(3):323.

Weinschenk, Aaron C, Isaac Erickson, Kaleigh Gale, Chloe Halford, James Harris, Alex Lange, Zea Miller, Ella Schwantes, Alax Stylinson, Connor Tenor et al. 2026. "General anxiety, political anxiety, and support for conspiracy theories." *Politics and the Life Sciences* 45(1):46–54.

Weinschenk, Aaron and Kevin Smith. 2024. "Is political anxiety different than general anxiety?" *Politics and the Life Sciences* 43(2):226–234.

Weiss, Chagai, Alexandra Siegel and Alexandra Scacco. Forthcoming. "Outgroup Avoidance." *Journal of Politics* .

Weiss, Chagai M. 2026. *Delivering Tolerance: How Unintended Institutional Inclusion Reduces Prejudice*. Princeton University Press.

Weiss, Chagai M, Donald P Green and Robb Willer. Forthcoming. "Can Politicians Bridge Partisan Divides? A Field Experiment with the US National Governors Association." *American Journal of Political Science* .

Weiss, Chagai M, Shira Ran and Eran Halperin. 2023. "Educating for inclusion: Diversity education programs can reduce prejudice toward outgroups in Israel." *Proceedings of the National Academy of Sciences* 120(16):e2218621120.

Westwood, Sean J, Justin Grimmer and Andrew B Hall. 2025. "Measuring perceived slant in large language models through user evaluations." *Stanford Business School* .

Yeomans, Michael, Julia Minson, Hanne Collins, Frances Chen and Francesca Gino. 2020. "Conversational receptiveness: Improving engagement with opposing views." *Organizational Behavior and Human Decision Processes* 160:131–148.

Young, Lauren E. 2019. "The psychology of state repression: Fear and dissent decisions in Zimbabwe." *American Political Science Review* 113(1):140–155.

Zeitzoff, Thomas. 2014. "Anger, Exposure to Violence, and Intragroup Conflict: A Lab in the Field Experiment in Southern Israel." *Political Psychology* 35(3):309–335.

Zeitzoff, Thomas. 2023. *Nasty politics: The logic of insults, threats, and incitement*. Oxford University Press.

Zhong, Wenjun, Jianghua Luo and Hong Zhang. 2024. "The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis." *Journal of affective disorders* 356:459–469.

Zilinsky, Jan and Thomas Zeitzoff. 2024. "Artificial Intelligence, Social Media, and the Politics of Anti-technology." Working paper, Abundance Institute. Accessed April 24, 2026.
URL: *https://abundance.institute/our-work/politics-of-anti-technology*

The Psychological Foundations of Liberal Democracy: Does Reducing Distress Increase Support for Democratic Norms and Out-Group Tolerance?

Supporting Information

Contents

A Complete questionnaire **1**

A.1 Screening Survey 1

A.2 Baseline and endline survey 4

A.3 Attention checks and exclusion rules 11

A.4 Engagement Survey 1 (E1)—onboarding 11

A.5 Engagement surveys 2–8 (E2–E8) 13

A.6 Placebo Engagement Survey 15

B Sub-scale aggregation and source instruments **16**

C Correlations among the H3b outcome measures **20**

D Power analysis **21**

E App installation and premium activation walkthrough, onboarding survey (E1) **25**

Statement of AI Use

Generative artificial intelligence (Claude Code Opus 4.7, 5.0, Fable) was used to assist with implementing and visualizing the simulation-based power analysis reported in Appendix D, generating the TikZ code for the study-flow diagram in Figure 1, language and copyediting, questionnaire development and response-scale harmonization, and literature search. All instances of AI use were supervised by the authors, and AI outputs were verified for their quality. We note that a key component of this study itself is based on generative AI (*Flourish*).

A Complete questionnaire

This appendix reproduces the full survey instrument. We field three instruments across the seven-week study: a screener (administered on day -3; Appendix A.1) and baseline/endline instruments (administered on Day 0 and Day 49; Appendix A.2). Treated respondents further engage with a sequence of eight brief engagement surveys E1–E8 administered to treated participants on Days 8, 12, 17, 24, 31, 38, 44, and 47, where E1 (Day 8) is the onboarding survey (Appendix A.4) and E2–E8 are the follow-up engagement surveys (Appendix A.5). In addition, all control-group respondents receive a brief placebo engagement survey, fielded at the same time as E2 (Day 12; Appendix A.6). Conditional on at least one Romano–Wolf-adjusted subscale rejecting in the H2 or H3 families at endline (Section 3 of the main text), we re-field the baseline/endline instrument as a follow-up survey approximately three months after endline (target Day 139) to all participants who completed the endline. The follow-up uses the same instrument as the endline and is therefore not separately reproduced below.

A.1 Screening Survey

A.1.1 Informed Consent

Study Contact. [Name redacted for peer review], [Institution redacted], [email redacted]
Institutional Contact. [Research Ethics Board redacted], [email redacted, phone redacted]

> "You are invited to participate in a brief online questionnaire for a research study exploring mental health, well-being, and perspectives on public life. Your responses are valuable, and there are no right or wrong answers. Please note that this survey does not offer medical diagnosis, treatment, or emergency care.
>
> **Eligibility.** To participate, you must be at least 18 years old, have been born in the United States, and possess a smartphone (iOS/Android) capable of installing an app. The survey can only be completed on an iOS or Android smartphone. This study is not intended for individuals experiencing an immediate mental health crisis or emergency.
>
> **What you will do.** After participating in a brief screener, you may be invited to complete a $\sim$15-minute survey about your background, mood and stress, social connection, and general social views. You may skip any question and stop at any time. Compensation is provided upon completing this survey in accordance with Forthright's terms. You may be invited to participate in a second module in which we will invite you to download and use an evidence-based well-being mobile app.
>
> **Voluntary participation and withdrawal.** Participation is voluntary. You may skip any question or withdraw without penalty and without losing compensation already earned.

Risks. If you choose to participate in this study, there is a medium risk of upset and psychological discomfort from survey questions and the use of a well-being application. As noted above, participation is voluntary, and you may choose to skip questions that you do not wish to answer.

Data privacy. Survey responses are stored with a unique study ID. Contact details remain with the panel provider and are not disclosed to the research team. All coded research data are securely stored, and results are reported in aggregate to maintain confidentiality. For scientific transparency, de-identified data and code may be shared. The research ethics program may have confidential access to data to verify that participant protection procedures are followed.

Conflict of Interest. Two members of the research team are employed by the provider of the mobile application under study and could benefit financially if the study findings portray the application in a positive light. These team members are also facilitating in-kind support for project implementation. To mitigate potential conflicts of interest in the study's implementation, data analysis, or reporting, these two team members will not be involved in implementation, data cleaning, analysis, or write-up. All analyses will follow the pre-registered analysis plan and will be conducted by other members of the research team.

For questions, contact [PI name redacted for peer review] at: [email redacted]. For participant rights, please contact the Research Ethics Board and reference protocol [REB contact redacted]."

Consent (check all to continue):

☐ I am 18 years or older.

☐ I have read and understood the information above.

☐ I voluntarily agree to participate.

☐ I agree to be re-contacted for follow-up surveys related to this project.

A.1.2 Screening

Please answer the following questions:

1. Willingness to Participate in Multiple Surveys. This study involves completing several brief online follow-up surveys over time. Are you willing and able to participate in multiple survey waves?

- Yes, I am willing and able
- No

2. Willingness to Participate in an App-Based Study. This study may involve downloading and regularly using a mobile app over a six-week period. The app has been downloaded by tens of thousands of users, many of whom reported enjoying their experience and benefiting from its use in various ways. In addition to the personal benefits you may gain from using the app, we plan to provide you with various rewards as compensation for the time you spend using the app.

Would you be willing to participate in a study that involves using an app for six weeks?

- Yes
- No

3. Feeling Thermometers—Democrats and Republicans. We'd like to get your feelings toward Democrats and Republicans on a "feeling thermometer." A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as

possible. You would rate the group at 50 degrees if you don't feel particularly positive or negative toward the group.

- Democrats
- Republicans
- *[Attention check; screening survey only]* To show you are reading carefully, please move this slider all the way to the right (100).

Response scale (0–100): 0 = coldest/most negative; 100 = warmest/most positive.

Note. The two substantive thermometer items (Democrats and Republicans) are administered here at the screener stage so that the affective-polarization eligibility check can be applied; the screening-survey measurements enter only this check. The third row of the battery is an instructed-response attention check (Appendix A.3); it is not a substantive measure and enters no subscale. The two thermometer items are re-administered at baseline and endline as part of the full feeling-thermometer battery (Appendix A.2.2), and the baseline-wave measurements serve as the pre-treatment values for the affective-polarization subscale.

At this stage, we screen on the following criteria:

- **Age ($\geq$ 18), born in the US, and Democrat/Republican identification** are verified by Forthright at the panel-profile level (variables `age`, `born_usa`, and `political_party_preference`) **prior to survey invitation**, so the survey is sent only to respondents already meeting these three criteria. The `political_party_preference` variable uses the standard 7-point branching item (1 = Strong Democrat, 2 = Democrat, 3 = Independent close to Democrat, 4 = Independent, 5 = Independent close to Republican, 6 = Republican, 7 = Strong Republican; plus "Other" and "No preference"); we invite only respondents who score 1, 2, 6, or 7. Independents (including leaners; codes 3–5), "Other," and "No preference" respondents are excluded prior to invitation. None of these variables is re-asked in the Qualtrics instrument. Forthright additionally supplies a panel-level `gender` variable for every respondent; this variable does not enter the eligibility screen but is used as a pre-registered subgroup covariate for the gender CATE (Section 4.4 of the main text).
- iOS or Android smartphone capable of installing *Flourish* (via Qualtrics)
- Willingness to participate in a longitudinal study (Q1)
- Willingness to participate in app-based study (Q2)
- Affective polarization (compute the in-party minus out-party feeling thermometer from Q3; retain only respondents whose value is **above the 25th percentile** of the corresponding distribution among partisans in the most recent ANES Time Series wave available at the time of the screening survey—i.e., retain the upper three quartiles. The specific threshold value will be inserted prior to fielding based on the latest ANES release.)
- Attention check (instructed-response row in Q3; correct response: 100). Pass rule and exclusion consequences in Appendix A.3.

A.2 Baseline and endline survey

A.2.1 Mental Health item battery

Current mental health or well-being support. (Asked only at baseline.) Are you currently receiving any mental health or well-being support? Please select all that apply.

- Individual therapy or counseling
- Group therapy or support group
- Psychiatric medication or medication for mental health
- Campus counseling or student support services
- Wellness coaching
- Mental health or well-being apps
- Other: _____

Depression (PHQ-8) and anxiety (GAD-7).

Over the last 2 weeks, how often have you been bothered by any of the following problems?

- **Depression**
  - Little interest or pleasure in doing things
  - Feeling down, depressed or hopeless
  - Trouble falling asleep, staying asleep, or sleeping too much
  - Feeling tired or having little energy
  - Poor appetite or overeating
  - Feeling bad about yourself—or that you're a failure or have let yourself or your family down
  - Trouble concentrating on things, such as reading the newspaper or watching television
  - Moving or speaking so slowly that other people could have noticed. Or, the opposite—being so fidgety or restless that you have been moving around a lot more than usual
- **Anxiety**
  - Feeling nervous, anxious or on edge
  - Not being able to stop or control worrying
  - Worrying too much about different things
  - Trouble relaxing
  - Being so restless that it is hard to sit still
  - Becoming easily annoyed or irritable
  - Feeling afraid as if something awful might happen

Response options: 0 = Not at all; 1 = Several days; 2 = More than half the days; 3 = Nearly every day.

Loneliness.

Please answer the following:

- How often do you feel that you lack companionship?
- How often do you feel left out?
- How often do you feel isolated from others?

Response options: Never; Almost never; Sometimes; Fairly often; Very often.

Perceived Stress.
In each case, indicate how often you felt or thought a certain way.

- In the last two weeks, how often have you felt that you were unable to control the important things in your life?
- In the last two weeks, how often have you felt confident about your ability to handle your personal problems?
- *[Attention check; baseline wave only]* To show you are reading carefully, please select "Almost never" for this item.
- In the last two weeks, how often have you felt that things were going your way?
- In the last two weeks, how often have you felt difficulties were piling up so high that you could not overcome them?

Response options: Never; Almost never; Sometimes; Fairly often; Very often.

Positive and negative affect (calm, well-being, vigor; depression, anxiety, anger). Please rate on a scale of 1 (not at all accurate) to 7 (extremely accurate) the extent to which each adjective accurately describes how you felt within the past two weeks.

- **Calm**
  1. Calm
  2. At ease
  3. Relaxed
- **Well-being**
  1. Happy
  2. Cheerful
  3. Pleased
- **Vigor**
  1. Full of pep
  2. Lively
  3. Energetic
- **Depression**
  1. Sad
  2. Unhappy
  3. Depressed
- **Anxiety**

1. On edge
2. Tense
3. Nervous

- **Anger**
  1. Hostile
  2. Angry
  3. Resentful

Scale: 1 = not at all accurate . . . 7 = extremely accurate.

Flourishing Scale. Below are eight statements with which you may agree or disagree. Using the 1–7 scale below, indicate your agreement with each item by indicating that response for each statement.

- I lead a purposeful and meaningful life.
- My social relationships are supportive and rewarding.
- I am engaged and interested in my daily activities.
- I actively contribute to the happiness and well-being of others.
- I am competent and capable in the activities that are important to me.
- I am a good person and live a good life.
- I am optimistic about my future.
- People respect me.

Response options: Strongly disagree; Disagree; Slightly disagree; Mixed or neither agree nor disagree; Slightly agree; Agree; Strongly agree.

A.2.2 Political Attitudes (H3)

Note. Items below are ordered to keep paired in-party / out-party items adjacent for respondent flow. Subscale-to-hypothesis-family assignment (H3a vs. H3b) is documented in Table 2 of the main text; the two out-party perception batteries below enter no H3 family and are exploratory secondary outcomes (Section 4.4 of the main text).

In-party norm violations.

Do you agree or disagree with the following statements:

- [inparty] elected officials should sometimes consider ignoring court decisions when the judges who issued those decisions were appointed by [outparty] presidents.
- If a [inparty] president can't get cooperation from [outparty] members of Congress to pass new laws, the [inparty] president should circumvent Congress and issue executive orders on their own to accomplish their priorities.
- *[Attention check; baseline wave only]* To show you are reading carefully, please select "Disagree" for this item.
- The government should be able to censor media sources that spend more time attacking [inparty] than [outparty].

Response options: Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree.

Out-party norm-compliance perception. Now, we would like you to give your best evaluation of how much [outparty] members agree with the following statements.

What percent of [outparty] voters do you think agree with the following statements?

- [outparty] elected officials should sometimes consider ignoring court decisions when the judges who issued those decisions were appointed by [inparty] presidents.
- If a [outparty] president can't get cooperation from [inparty] members of Congress to pass new laws, the [outparty] president should issue executive orders on their own to accomplish their priorities.
- The government should be able to censor media sources that spend more time attacking [outparty] than [inparty].

Response scale (0–100): 0 = 0% of out-party voters agree; 100 = 100% agree.

Opposition to political violence.

Please consider the following scenario:

> "A person was convicted of vandalism. He was arrested by police after he vandalized several large signs expressing support for candidates of the [outparty] party."

- Do you support or oppose the person's actions?

Response options: Strongly support; Support; Neither support nor oppose; Oppose; Strongly oppose.

Perceived out-party support for political violence.

Now, please consider another scenario:

> "A person was convicted of vandalism. He was arrested by police after he vandalized several large signs expressing support for candidates of the [inparty] party."

- What percent of [outparty] voters do you think support the person's actions?

Response scale (0–100): 0 = 0% of out-party voters support the action; 100 = 100% support.

Support for Bipartisan Cooperation.

- To what extent would you like to see Democratic and Republican elected representatives work together?

Response scale (0–100): 0 = not at all; 100 = very much.

Generalized trust.

- On a scale from 0 to 10, where 0 is not at all and 10 is completely, in general, how much do you trust most people?

Response scale (0–10): 0 = not at all (. . .) 10 = completely.

Feeling thermometers. We'd like to get your feelings toward a number of groups on a "feeling thermometer." A rating of zero degrees means you feel as cold and negative as possible. A rating of 100 degrees means you feel as warm and positive as possible. You would rate the group at 50 degrees if you don't feel particularly positive or negative toward the group.

- Democrats
- Republicans
- LGBTQ people
- Undocumented immigrants in the US
- Ivy-league academics
- Wealthy Americans (top 1%)
- Gun owners
- Evangelical Christians

Response scale (0–100): 0 = coldest/most negative; 100 = warmest/most positive.

All eight thermometer items are administered to every respondent at both baseline and endline; the Democrats and Republicans items are additionally administered in the screening survey for the affective-polarization eligibility check (Appendix A.1). The partisan-conditional H3b out-group-thermometer subscale is constructed from the three items that are out-group for the respondent's party—for Republicans: LGBTQ people, undocumented immigrants in the US, and Ivy-league academics; for Democrats: wealthy Americans (top 1%), gun owners, and evangelical Christians. These six items were selected to capture out-group identities commonly associated with the opposing party's coalition. The remaining three out-group items per respondent are collected and stored, but do not enter any pre-registered H3 subscale and are reported only as exploratory descriptive statistics. **Secondary outcomes.**

Political interest and turnout intention.

- How likely are you to vote in the next national elections? (0–100)
- How interested are you in politics?

Political interest response options: Very interested = 4; Quite interested = 3; Hardly interested = 2; Not at all interested = 1. Matches the European Social Survey core module (structurally equivalent to WVS Q199) and Ojeda, Bernardi and Landwehr (2023). Coded so higher values correspond to more interest.

A.2.3 Intermediate Psychological Orientations (H2)

Zero-Sum Thinking.

Do you agree or disagree with the following statements:

- **Ethnic:** In the United States, there are many different ethnic groups (Blacks, whites, Asians, Hispanics, etc.). If one ethnic group becomes richer, this generally comes at the expense of other groups in the country.
- **Citizenship:** In the United States, there are those with American citizenship and those without. If those without American citizenship do better economically, this will generally come at

the expense of American citizens.

- **Trade:** In international trade, if one country makes more money, then it is generally the case that the other country makes less money.
- **Income:** In the United States, there are many different income classes. If one group becomes wealthier, it is usually the case that this comes at the expense of other groups.

Response options: Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree.

Conversational Receptiveness.

Do you agree or disagree with the following statements:

- I am generally curious to find out why other people have different opinions than I do.
- Listening to people with views that strongly oppose mine tends to make me angry.

Response options: Strongly disagree; Somewhat disagree; Slightly disagree; Neither agree nor disagree; Slightly agree; Somewhat agree; Strongly agree.

Political Anxiety. "How much anxiety does each of the following give you?" (shortened)

1. The election of a disliked candidate or political party
2. The level of polarization and conflict in the current political climate
3. That the American public is insufficiently informed about politics
4. The poor quality of political leaders/candidates

Response scale (1–10): 1 = no anxiety at all . . . 10 = a great deal of anxiety.

Pessimism.

Do you agree or disagree with the following statements:

- In uncertain times, I usually expect the best.
- Overall, I expect more good things to happen to me than bad.
- I hardly ever expect things to go my way.
- I rarely count on good things happening to me.

Response options: Strongly disagree; Disagree; Neither agree nor disagree; Agree; Strongly agree.

Source: Items adapted from the Life Orientation Test–Revised (LOT-R; Scheier, Carver and Bridges 1994); the two **positively-worded** items ("I usually expect the best", "I expect more good things than bad") are reverse-coded so that higher subscale scores correspond to greater pessimism.

A.2.4 Attitudes Toward AI

To what extent do you agree with the following statements? [101-point scale]

- The benefits of artificial intelligence (AI) outweigh the risks for me personally.
- The benefits of artificial intelligence (AI) outweigh the risks for my community.
- The benefits of artificial intelligence (AI) outweigh the risks for society as a whole.

To what extent would you support or oppose the use of AI systems for the following tasks? [101-point scale, order randomized]

- Diagnosing diseases from medical scans.
- Helping judges determine prison sentences.
- Screening job applicants for interviews.
- Deciding who is approved for a bank loan.
- Identifying potential criminal suspects from security footage.
- Providing mental health support through a chatbot.

A.2.5 Endline Diagnostic

Before you finish, we have a few more questions. In your own words, please tell us what you think we are focused on in this study. [Include an explicit "Don't Know box."]

A.2.6 AI Use and Views About AI

Note. The open-ended item is not asked at baseline. The two use items are asked at baseline and endline. At endline, all three items are displayed after the Endline Diagnostic, so that self-reported use of AI products cannot inform it; at baseline, the use items directly follow the Attitudes Toward AI battery. Rows are displayed in a fixed order. None of these items enters a pre-specified outcome of this report; like the Attitudes Toward AI battery, they are reported in the companion paper.

Views about AI (open-ended). (Not asked at baseline.)
In your own words, what do you think about artificial intelligence (AI), and why? Please write a few sentences. There are no right or wrong answers. [Open text box]

Use of AI products.
In the past two weeks, on how many days did you use each of the following types of AI products?

- General-purpose AI chatbots or assistants (for example, ChatGPT, Gemini, Claude, Copilot, or Meta AI)
- AI tools for creating images, videos, or music (for example, Midjourney, Sora, Suno, or the image tools in ChatGPT or Gemini)
- Mental health or well-being apps with a built-in AI coach or chatbot (for stress, mood, sleep, or emotional support)
- AI companion apps for conversation, friendship, or role-play (for example, Replika or Character.AI)

Response options: Not at all; 1–2 days; 3–6 days; 7–10 days; 11–14 days.

General-purpose AI use for emotional support.

- In the past two weeks, on how many days did you use a general-purpose AI chatbot or assistant (for example, ChatGPT, Gemini, Claude, Copilot, or Meta AI) for emotional support, or to talk through stress, your mood, or personal problems? Please do not count apps made specifically for mental health or well-being.

Response options: Not at all; 1–2 days; 3–6 days; 7–10 days; 11–14 days.

A.3 Attention checks and exclusion rules

The study embeds three attention checks, all administered before treatment assignment. The screening survey (Appendix A.1) contains one instructed-response row in the feeling-thermometer battery (Q3), which asks respondents to move the slider all the way to the right; the correct response is exactly 100. We apply no tolerance band: 100 is an endpoint of the scale, so a drag that overshoots is clamped to the maximum, and the rule is therefore unambiguous to obey on the smartphone-only instrument and mechanical to score. The baseline instrument (Appendix A.2) contains two instructed-response items, at baseline (pre-treatment) only: (i) one in the Perceived Stress battery (correct answer: "Almost never"), and (ii) one in the In-party norm violations battery (correct answer: "Disagree").

Respondents who fail the screening-survey check are ineligible and are not invited to the baseline survey. Respondents who fail either baseline check are screened out prior to treatment assignment and are not part of the randomized analytic sample. None of the three items enters any subscale.

A.4 Engagement Survey 1 (E1)—onboarding

E1 is the onboarding survey: it is the first of eight engagement surveys (E1–E8) and is delivered on Day 8 (eight days after the Day-0 baseline) to treated participants only—the randomly selected 50% of the enrolled sample (respondents who completed the baseline survey and passed both baseline attention checks).

Boost Your Well-being: Free three-month App Access + Bonus Rewards.

> "Congratulations! You have been selected to take part in a research study where participants receive free premium access to an award-winning well-being app. As part of this study, you will get a completely free three-month subscription to *Flourish*, an AI-powered well-being application. This premium service normally costs $12.99 per month, but you will receive full access at no cost.
>
> As a participant in this study, you are invited to make regular use of your free premium access to the *Flourish* app. To support and recognize your continued engagement, you are eligible for the following rewards:"

- **Free three-month premium access to *Flourish*:** you will receive a three-month subscription to *Flourish*, a leading AI-powered well-being application, valued at $12.99 per month.
- **Weekly lotteries ($3,000 total prize pool):** during the six-week study, you accumulate lottery tickets in proportion to your number of "flourished days" on the in-app counter. The more days you use the app, the greater your chance of winning. For each week of the study, five participants are randomly drawn to receive $100 each, for a total of 30 winners and $3,000 in

prizes across the six-week study. All draws are conducted, and winners notified, when these rewards are paid out.

- **$20 guaranteed bonus:** You will automatically receive a $20 bonus if you reach at least 20 "flourished days" on the in-app counter over the six-week study period.
- **Paid engagement surveys:** this is the first of eight brief, paid engagement surveys (E1–E8) that you will receive over the six-week study; the remaining seven (E2–E8) each briefly remind you of the rewards above, offer a short tip on getting the most out of *Flourish*, and deep-link into three activities within the *Flourish* app, of which you choose one to complete. Compensation for each engagement survey is contingent on completing one of the three linked in-app activities and, together with the lottery and flourished-days bonus, is disbursed within four weeks after the end of the six-week study.

"Importantly, in this study, we will not collect any app-related data other than your overall usage (for example, the days and time you spend in the app, and activities you complete). All conversations and activities within the app will remain completely private; researchers will not have access to your in-app content.

Please note that participation is voluntary, and this app is not a form of medical care. If you feel distressed, you may stop at any time."

Consent.

☐ I am 18+ and agree to take part in this study.

App installation, premium activation, and Buddy ID.

To install the Flourish app, activate premium access, and link their app account to the study, participants are guided through an in-survey walkthrough (with device-specific download instructions for iOS and Android). Participants (i) install the app and create an account; (ii) during the app's onboarding, select "I have an organization code" and enter the study-specific code displayed in the survey, which activates premium access for the three-month study period; and (iii) retrieve their six-character Flourish Buddy ID from the app's Community tab and enter it in the survey. The full walkthrough, including all in-app screenshots, is reproduced in Appendix E.

- Please enter your *Flourish* Buddy ID: _____

As a final step, please confirm each of the following:

☐ The *Flourish* app is installed on my phone and I have created an account.

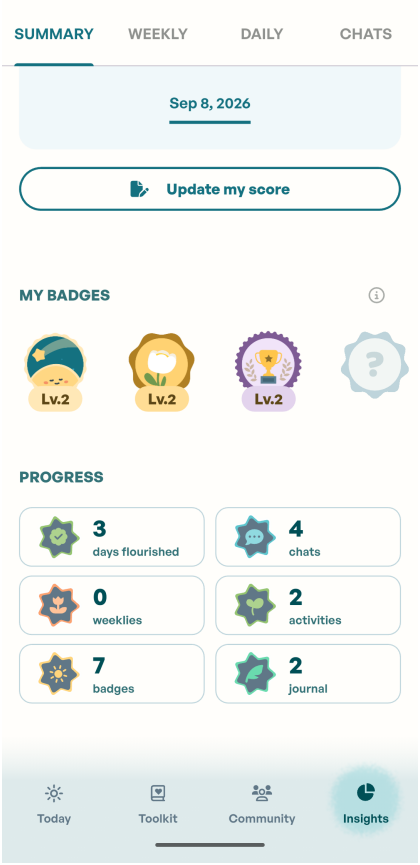
☐ I entered the organization code and my premium access is active.

☐ I entered my *Flourish* Buddy ID.

"Over the next six weeks, please aim to use the app regularly; each day you complete at least one in-app activity counts as a "flourished day" on the in-app counter and increases both your chance of winning a weekly lottery prize and your progress toward the $20 bonus. It might feel a bit awkward at first. But our extensive experience suggests that *Flourish* users love the app. So give it a shot!

You can check your number of flourished days at any time in the app: tap *Insights* in the bottom menu bar and look for the "days flourished" tile under *Progress*. [Screenshot of the Insights page displayed here; reproduced as Figure A.1.]"

Figure A.1: Insights page of the *Flourish* app, as shown in the onboarding survey (E1).



Note: Screenshot of the Summary view of the Insights tab in the *Flourish* app, displayed to treated respondents in the onboarding survey (E1). The "days flourished" tile under Progress is the in-app counter on which the weekly lotteries and the $20 bonus are based.

> "Please stay tuned for the remaining seven brief engagement surveys (E2–E8), which will be distributed via Forthright over the six-week study period."

A.5 Engagement surveys 2–8 (E2–E8)

Sent to treated participants only, on Days 12, 17, 24, 31, 38, 44, and 47 (E2–E8). Together with E1 (the onboarding survey on Day 8; Appendix A.4), the engagement-survey schedule places two surveys in the first treatment week, one in each of treatment weeks 2–5, and two in treatment week 6. Advancement to E2 is conditional on completing E1, that is, on reporting a Flourish Buddy ID that Flourish matches to an account with active premium access; a treated respondent who reaches the E2 fielding window without having completed E1, or whose reported Buddy ID cannot be matched, is re-served E1 (the onboarding survey) in place of E2. Once a respondent has cleared the E1 gate, E3–E8 are fielded on the original schedule regardless of completion of any individual subsequent wave. Respondents who never complete onboarding continue to receive E1 in every fielding window until the treatment window closes.

Each engagement survey shows the participant three short components, reproduced below.

1. Reminder of incentive structure.

"Welcome back, and thank you for continuing to take part in this study. As a reminder, during the six-week study, you can earn:

- **Weekly lotteries ($3,000 total prize pool):** you accumulate lottery tickets in proportion to your number of "flourished days" on the in-app counter. For each week of the study, five participants are randomly drawn to receive $100 each, for a total of 30 winners and $3,000 in prizes across the six-week study. All draws are conducted, and winners notified, when these rewards are paid out.
- **$20 guaranteed bonus:** if you reach at least 20 "flourished days" on the in-app counter over the six-week study, you will automatically receive a $20 bonus.
- Compensation for engagement surveys like this one, contingent on completing one of the three linked in-app activities. Together with the lottery and flourished-days bonus, this compensation is disbursed within four weeks after the end of the six-week study.

2. Short guidance on effective app use.

"To help you get the most out of *Flourish*, here is a short tip: [brief guidance, e.g., "Set aside a quiet 10 minutes so you can focus without interruptions," "Be open to trying different app components; each one exercises a different skill," or "Return to the same activity on multiple days; most *Flourish* exercises deepen with repetition."]"

3. Deep links to today's activities.

"Please tap one of the three links below to open *Flourish* and complete the activity of your choice. Payment for this engagement survey is contingent on completing one of the three linked in-app activities and will be disbursed, together with the lottery and flourished-days bonus, within four weeks after the end of the six-week study.

`flourish://activity/[activity_id_1]`
`flourish://activity/[activity_id_2]`
`flourish://activity/[activity_id_3]`"

Note. Each of the engagement surveys deep-links to three in-app activities within *Flourish*, of which the participant chooses one to complete; the set of three activities rotates across waves. Activities span the full range of in-app components (conversations with *Sunnie*, interactive journaling, and science-based Toolkit activities, e.g., guided breathing and skill-practice modules); the assignment of activities to engagement-survey waves is pre-specified. Completion of a linked in-app activity is verified through privacy-preserving usage metadata that matches participant identity to app activity through the *Flourish* Buddy ID reported in E1, and compensation is contingent on verified completion of one of the three. Each engagement survey contains a reminder that *Flourish* is a wellness application, not a substitute for clinical care, and that respondents who feel distressed may stop at any time.

A.5.1 Experience With the Flourish App

Note. Asked only at the end of E8 (Day 47). Both items enter no analysis in this report and are reported in the companion paper.

Finally, two questions about the Flourish app that you were invited to use as part of this study.

- Overall, how helpful did you find the *Flourish* app?

Response options: Not at all helpful; Slightly helpful; Moderately helpful; Very helpful; Extremely helpful.

- Has using the *Flourish* app changed your views about artificial intelligence (AI)—for example, what AI can do, or its benefits and risks?

Response options: Much more negative; Somewhat more negative; No change; Somewhat more positive; Much more positive.

A.6 Placebo Engagement Survey

We recently asked you to participate in a survey related to social issues. As a token of appreciation, we are including you in a lottery in which five participants will be randomly drawn to receive $100 each. The lottery will be implemented in a single draw on the following date: [predetermined date after the endline survey; inserted prior to fielding]. If you are selected, we will contact you, and you will receive a $100 prize from Forthright.

B Sub-scale aggregation and source instruments

This section sets out the aggregation rule for our primary subscale outcomes.[18]

Aggregation rule. Let y_{ij} denote respondent i's response to item j within subscale s, after reverse coding items so that higher values correspond to the subscale's named direction. Because items within a subscale occasionally use different response scales (for example, the 4-option political-interest item and the 0–100 turnout-intention item within the political-interest-and-turnout-intention subscale), we first z-standardize each item on the baseline distribution, placing items on a common scale before aggregation:

$$\tilde{y}_{ij} = \frac{y_{ij} - \bar{y}_j^0}{\sigma_j^0},$$

where $\bar{y}_j^0$ and σ_j^0 are the mean and standard deviation of item j on the *baseline wave, pooled across both arms, computed on observed (non-missing) responses*. The same pooled-baseline mean $\bar{y}_j^0$ is used for baseline imputation below.

Item-level missingness. We handle item-level missingness asymmetrically across waves. *At baseline*, we impute any missing item value y_{ij} with the baseline cross-respondent mean $\bar{y}_j^0$ before z-scoring; the imputed item then z-scores to exactly 0, the cross-respondent average on the standardized scale. After imputation, every respondent has a non-missing value on every baseline item, so Y_{ib}^{pre} in Equation 1 is non-missing for every respondent and every subscale. *At endline*, we do not impute: missing items remain missing and do not contribute to the subscale.

Subscale construction. The subscale score Y_{is} is the equal-weighted mean of the item z-scores,

$$Y_{is} = \frac{1}{|\mathscr{J}_i|} \sum_{j \in \mathscr{J}_i} \tilde{y}_{ij},$$

where $\mathscr{J}_i \subseteq \{1, \ldots, n_s\}$ is the set of items contributing to respondent i's subscale. At baseline, $\mathscr{J}_i = \{1, \ldots, n_s\}$ for every respondent (after the imputation step above). At endline, $\mathscr{J}_i$ is the set of items with a non-missing response, and Y_{is} is treated as missing only if all n_s items are missing ($|\mathscr{J}_i| = 0$). We then re-z-standardize Y_{is} on the same baseline distribution, $\tilde{Y}_{is} = (Y_{is} - \bar{Y}_s^0)/\sigma_s^0$, so that the subscale-level ITT in Equation 1 is expressed in baseline-SD units. Each $\tilde{Y}_{is}$ enters Equation 1 directly as a primary outcome; we do not aggregate across subscales. Single-item subscales reduce to $\tilde{Y}_{is} = \tilde{y}_{i,1}$.

Response-scale harmonization and reverse coding. Several subscales use a response scale that differs from the source instrument, including cases where the number of response options differs (e.g., SAS 0–4 $\rightarrow$ 1–7). These harmonizations do not threaten the subscale-level ITT estimates in Equation 1: every respondent in both arms answers the identical harmonized format at both waves, and each item enters the subscale after item-level z-standardization on the baseline distribution,

[18] The affect, loneliness, and flourishing subscales follow the trial reported in Cachia et al. (2026a), and we retain the original item-to-subscale assignments; the remaining H1 subscales are taken directly from their source instruments (Table B.1). Relative to Cachia et al., we harmonize response scales across surveys and aggregate items as set out below. Table B.1 maps each subscale to its source instrument.

so all treatment–control contrasts are estimated on a common within-sample scale expressed in baseline-SD units. We reverse-code items where needed so that each subscale has a consistent directional interpretation.

Mental health and well-being (H1 primary subscales). PHQ-8 and GAD-7 retain the source 0–3 scale and two-week reference window (Kroenke et al., 2009; Spitzer et al., 2006); we drop PHQ-9 item 9 (suicidal ideation) per the PHQ-8 convention. We expand UCLA-3 loneliness items from the source 3-option scale (hardly ever / some of the time / often; Hughes et al. 2004; also used by Cachia et al. 2026a) to a 5-option scale (never / almost never / sometimes / fairly often / very often) to harmonize the response options with PSS-4. PSS-4 items use 1–5 (Cohen, Kamarck and Mermelstein 1983; Cohen 1988: 0–4); we reverse-code items 2 ("felt confident…") and 3 ("things going your way"); item indices refer to scored items only, excluding the embedded attention check at baseline. We shorten the PSS reference window from "in the last month" to "in the last two weeks" to match the PHQ-8/GAD-7 window within the same mental-health battery. SAS negative-affect items use 1–7 (Cachia et al. 2026a: 0–4); item assignment to the three sub-components (Depression, Anxiety, Anger) matches Cachia et al. (2026a, Table S3). Flourishing scale retains 1–7 (all eight items). SAS positive-affect items use 1–7 (Cachia et al. 2026a: 0–4); item assignment to the three sub-components (Calm, Well-being, Vigor) matches Cachia et al. (2026a, Table S3).

Intermediate psychological orientations (H2 primary subscales). Zero-sum thinking items use the source 5-point Likert scale (Strongly disagree… Strongly agree). Conversational-receptiveness items use the source 7-point Likert scale (Strongly disagree… Strongly agree); we field two of the eighteen dispositional-scale items (source items 5 and 15), presented without the scale's issue-framing preamble; we reverse-code item 2 ("listening… makes me angry") so higher = more receptive. Political-anxiety items use the source 1–10 scale (1 = no anxiety; 10 = a great deal); we shorten the battery from eight to four items (items 1, 2, 3, and 6 in the ordering of Smith, Weinschenk and Panagopoulos 2024). Pessimism items use a 5-point Likert scale (Strongly disagree… Strongly agree) and are drawn from the Life Orientation Test–Revised (LOT-R; Scheier, Carver and Bridges 1994), narrowed from six scored items to four (two positively-worded and two negatively-worded items); we reverse-code the two positively-worded items so that higher subscale scores correspond to greater pessimism.

Political attitudes (H3 primary subscales). H3a (support for democratic norms). PRL Pulse in-party norm-violation items use the source 5-point Likert scale (Strongly disagree… Strongly agree); we field three of the six PRL norm items (`norm_judges`, `norm_executive`, `norm_censorship`) and their three perception counterparts; higher subscale values reflect greater endorsement of in-party norm violations (predicted negative ITT). The opposition-to-political-violence subscale is a single item drawn from the PRL Pulse political-violence vignette battery (a person convicted of vandalism for vandalizing several large signs supporting out-party candidates), measured on a 5-point Likert scale (Strongly support… Strongly oppose), reverse-coded so higher = stronger opposition to violence. We substitute "a person" for PRL's permuted personal name in the vignette stem to avoid name-based confounds (e.g., race/ethnicity signaling from the chosen name); all other vignette wording matches Polarization Research Lab (2024) verbatim. Unlike the original PRL Pulse fielding, in which this vignette (`violence2`) is administered only to respondents who endorse the lower-severity vignette (`violence1`) earlier in the cascade, we administer it stand-alone to all respondents. The bipartisan-cooperation item uses 0–100, coded so higher = greater support for cooperation. *H3b (attitudes toward out-groups).* Generalized trust uses 0–10 (single item; predicted

positive ITT). Feeling thermometers use 0–100. All eight thermometers (Democrats, Republicans, LGBTQ people, undocumented immigrants in the US, Ivy-league academics, wealthy Americans (top 1%), gun owners, evangelical Christians) are administered to every respondent at both baseline and endline; the Democrats and Republicans thermometers are additionally administered in the screening survey (day -3), where they enter only the affective-polarization eligibility check; the screening-survey battery also contains a non-substantive instructed-response attention-check row (Appendix A.3) that enters no subscale. The eight thermometers enter as two H3b subscales: (i) the partisan-conditional out-group thermometer, constructed for each respondent as the equal-weighted mean of three z-scored out-group thermometers selected by the respondent's party—for Republicans: LGBTQ people, undocumented immigrants in the US, and Ivy-league academics; for Democrats: wealthy Americans (top 1%), gun owners, and evangelical Christians (predicted positive ITT); and (ii) affective polarization, defined as the respondent's in-party thermometer minus out-party thermometer (range -100 to $+100$, predicted negative ITT). The affective-polarization subscale is well-defined for the entire analytic sample, since eligibility (Section 3.1) is restricted to respondents who identify as Democrat or Republican. The remaining three out-group items per respondent are stored but do not enter any pre-registered subscale or Romano–Wolf correction; they may be reported as exploratory descriptive statistics only.

Secondary outcomes. Three prespecified exploratory secondary outcomes are reported with unadjusted two-sided p-values and are not part of the Romano–Wolf correction. (i) Political interest and turnout intention: the political-interest item ("How interested are you in politics?") uses the 4-option ESS core-module scale (not at all / hardly / quite / very interested; Ojeda, Bernardi and Landwehr 2023), coded so higher = more interest; the turnout-intention item uses 0–100. (ii)–(iii) Out-party perceptions (PRL Pulse battery; Polarization Research Lab 2024), administered at both baseline and endline and aggregated with the same two-step rule as the primary subscales: the out-party norm-compliance perception items use 0–100 (what percent of out-party voters agree), reverse-coded so higher = greater perceived out-party compliance; the perceived-out-party-support-for-political-violence subscale is a single 0–100 item asking what percent of out-party voters support the action in a mirrored version of the H3a vandalism vignette, in which the vandalized signs support in-party candidates; higher values reflect greater perceived out-party support for violence. The original PRL Pulse battery does not field a perceived-out-party-support item for the vandalism vignette (`violence2`); we adapt PRL's perceived-support item format to this vignette accordingly. All three secondary outcomes appear in Table B.1. The attitudes-toward-AI battery (Appendix A.2.4), the AI-use and open-ended items (Appendix A.2.6), and the two items on treated respondents' experience with *Flourish* (Appendix A.5.1) are collected but do not enter any pre-specified outcome.

Table B.1: Primary and secondary subscales and source instruments.

Subscale (ours)	Source instrument
Mental health and well-being (H1)	
PHQ-8 depression	PHQ-8 (Kroenke et al., 2009)
GAD-7 anxiety	GAD-7 (Spitzer et al., 2006)
UCLA loneliness	UCLA-3 (Hughes et al., 2004)
Perceived stress	PSS-4 (Cohen, Kamarck and Mermelstein, 1983; Cohen, 1988)
Negative affect (depression, anxiety, anger)	Subcomponents of Affect Scale (Jenkins et al., 2023); item assignment as in Cachia et al. (2026a)
Flourishing	Flourishing Scale (Diener et al., 2010)
Positive affect (calm, well-being, vigor)	Subcomponents of Affect Scale (Jenkins et al., 2023); item assignment as in Cachia et al. (2026a)
Intermediate psychological orientations (H2)	
Zero-sum thinking	Chinoy et al. battery (Chinoy et al., 2026)
Conversational receptiveness	Yeomans et al. (2020)
Political anxiety	Political anxiety scale (Smith, Weinschenk and Panagopoulos, 2024)
Pessimism	Life Orientation Test–Revised (LOT-R) items (Scheier, Carver and Bridges, 1994)
Support for democratic norms (H3a)	
In-party norm violations	PRL Pulse battery (Polarization Research Lab, 2024)
Opposition to political violence	PRL Pulse battery (Polarization Research Lab, 2024)
Support for bipartisan cooperation	Megastudy item (Voelkel et al., 2024)
Attitudes toward out-groups (H3b)	
Generalized trust	Single 0–10 generalized-trust item
Out-group thermometer (partisan-conditional)	Standard 0–100 feeling thermometer (cf. Pew Research Center 2017)
Affective polarization (in-party − out-party thermometer)	Standard 0–100 feeling thermometer (cf. Pew Research Center 2017)
Secondary outcomes	
Political interest and turnout intention	ESS core-module interest item (as used by Ojeda, Bernardi and Landwehr 2023); single turnout-intention item
Out-party norm-compliance perception	PRL Pulse battery (Polarization Research Lab, 2024)
Perceived out-party support for political violence	Adapted from PRL Pulse battery (Polarization Research Lab, 2024)

C Correlations among the H3b outcome measures

We use the 2024 wave of the American National Election Studies (ANES) Time Series study (American National Election Studies, 2026) to examine the empirical association among the three subscales of the H3b family. We restrict the ANES sample to respondents who identify as Democrats or Republicans on the root party-identification item ($N = 3{,}517$), mirroring the partisanship eligibility criterion of our design (Section 3.1 of the main text).

We construct an analog of each subscale from the ANES items. Generalized trust is the item "Generally speaking, how often can you trust other people?", reverse-coded so that higher values indicate more trust; the ANES version uses a five-point frequency scale rather than our 0–10 scale. Affective polarization is the in-party minus out-party feeling thermometer. For the out-group thermometer, we mirror the aggregation rule in Appendix B: each 0–100 thermometer is z-scored on the partisan analysis sample, and we average the thermometers toward the closest ANES analogs of our target groups—for Democrats, big business (wealthy Americans), the National Rifle Association (gun owners), and Christian fundamentalists (evangelical Christians); for Republicans, gay men and lesbians and transgender people (averaged into a single LGBTQ proxy) and illegal immigrants (undocumented immigrants). The 2024 wave contains no analog for Ivy League academics. All estimates are weighted with the ANES full-sample post-election weight.

Table C.1 reports pairwise weighted Pearson correlations. The two thermometer-based subscales are weakly correlated: respondents who rate their respective out-groups more warmly are less affectively polarized ($r = -0.28$). Generalized trust, by contrast, is not meaningfully correlated with the two partisan measures ($r = -0.06$ and -0.02). We therefore treat generalized trust as capturing a related but distinct dimension of out-group orientation.

Table C.1: Correlations among analogs of the H3b subscales in the ANES 2024 Time Series.

	(1)	(2)	(3)
(1) Generalized trust	—		
(2) Out-group thermometer (partisan-conditional)	−0.06	—	
(3) Affective polarization (in-party − out-party)	−0.02	−0.28	—

Note: Pairwise weighted Pearson correlations among partisans (root-party-ID Democrats and Republicans) in the ANES 2024 Time Series, weighted with the full-sample post-election weight (variable V240107b); pairwise-complete observations ($N = 3{,}176$, 3,188, and 3,172 for the (1)–(2), (1)–(3), and (2)–(3) pairs, respectively). Generalized trust and the party thermometers are pre-election items; the group thermometers are post-election items. The out-group thermometer analog omits the Ivy League academics target group, which has no ANES 2024 analog.

D Power analysis

This appendix provides the full data-generating process, design parameters, and sensitivity analyses behind the headline power numbers reported in Section 4.2 of the main text. We compute power by simulation in the *DeclareDesign* framework (Blair et al., 2019), declaring the data-generating process, treatment assignment, and estimator as separate steps. The model draws baseline outcomes $Y_i^{\text{pre}} \sim \mathcal{N}(0,1)$ and endline outcomes $Y_i^{\text{post}} = \rho\, Y_i^{\text{pre}} + (\tau^{\text{ITT}}/c)\, D_i Z_i + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 1-\rho^2)$ and $D_i \sim \text{Bernoulli}(c)$ is a per-respondent compliance indicator at $c = 0.65$; non-compliers in the treated arm receive no treatment effect, and the per-complier per-subscale effect is set to τ^{ITT}/c so that the population ITT equals τ^{ITT} in expectation while compliance heterogeneity inflates within-treated variance. Treatment is block-randomized within the twenty blocks, and 10% of respondents are dropped from estimation (MCAR). We estimate the per-subscale ITT by the OLS specification in Equation 1 with HC1 standard errors, tested one-sided at $\alpha = 0.05$ in the pre-registered direction. For the family-wise analysis, we extend the model to a K-outcome design in which the K subscales come from a latent-factor decomposition with within-subscale baseline–endline correlation $\rho = 0.8$ and within-family between-subscale correlation $\rho_{\text{sub}} = 0.5$, and apply the Romano–Wolf step-down to the resulting K-vector of test statistics under the alternative ($\tau = \tau^{\text{ITT}}$) and the null ($\tau = 0$).

For this power analysis, we operationalize compliance as a binary onboarding indicator: a treated respondent is a complier if they complete the onboarding survey—install *Flourish*, activate premium access with the study code, and report a *Flourish* Buddy ID that matches an account on the *Flourish* back-end. We assume that 65% of treated respondents complete onboarding; conditional on onboarding, we expect high cumulative in-app usage given the engagement-survey incentive architecture (Section 3.3 of the main text).

Table D.1 reports the design parameters and headline power numbers; Figure 3 of the main text plots power as a function of the true per-subscale ITT. At $N_{\text{eff}} = 2{,}430$, $\rho = 0.8$, 65% treated-arm compliance, and one-sided $\alpha = 0.05$ prior to correction, the trial has 99.3% analytical power to detect an ITT of 0.10 SD on any single subscale; 5,000 simulations of the full data-generating process likewise give 99.3% power. The minimum detectable effect per subscale at 95% power is 0.080 SD. Under the Romano–Wolf step-down at $\rho_{\text{sub}} = 0.5$, family-wise power at the design ITT is 99.9% for H1 (seven subscales), 99.9% for H2 (four subscales), 99.9% for H3a (three subscales), and 99.9% for H3b (three subscales), with an expected 6.90 of 7 H1 subscales rejecting, 3.95 of 4 for H2, 2.97 of 3 for H3a, and 2.97 of 3 for H3b. Reassuringly, the same step-down evaluated under $\tau = 0$ rejects in 4.0% to 5.4% of simulations across the four families: no family exceeds the targeted 5% family-wise error rate by more than Monte Carlo noise.

Figure D.1 shows how per-subscale power varies as a function of the baseline–endline correlation ρ, holding all other design parameters fixed at the values in Table D.1. The trial remains well powered even for relatively small baseline–endline correlations: per-subscale power exceeds 95% for $\rho \geq 0.66$.

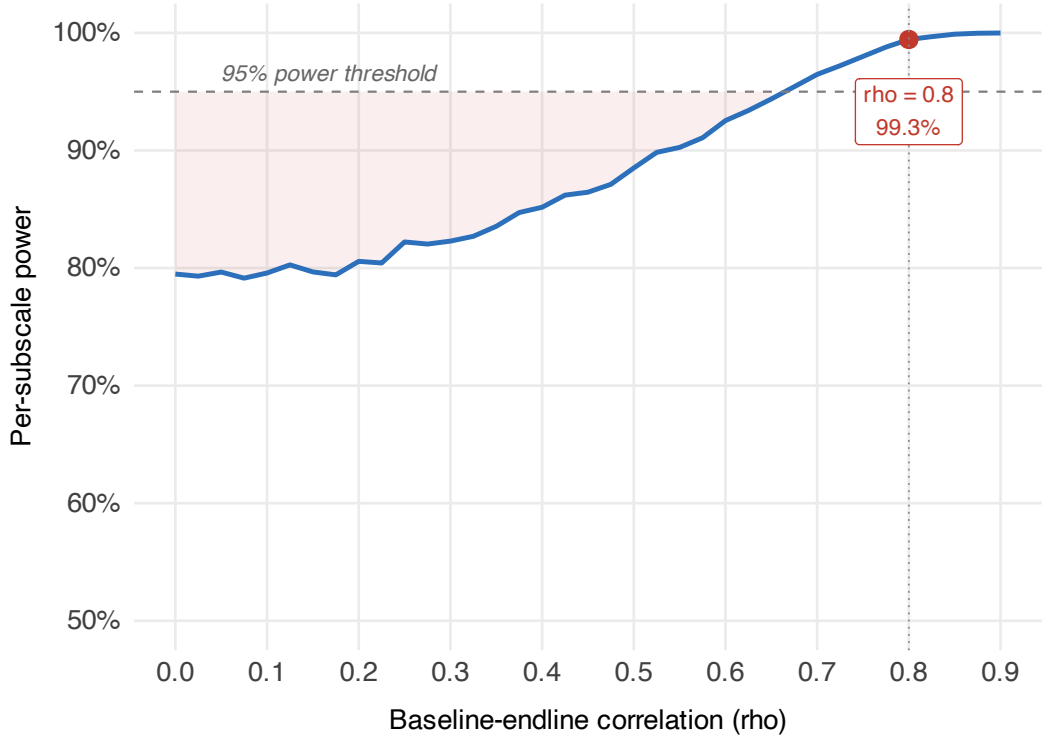
Figure D.2 repeats the family-wise analysis as a function of the within-family between-subscale correlation ρ_{sub}, for which we lack a strong prior. Family-wise power under the Romano–Wolf step-down is insensitive to this parameter: at the design ITT, power remains above 99% in all four families across $\rho_{\text{sub}} \in [0, 0.9]$.

Table D.1: Design parameters for the power analysis.

Parameter	Value
Sample and assignment	
Total baseline sample (N)	2,700
Treatment allocation (treated / control)	50% / 50%
Attrition (MCAR)	10%
Effective sample (N_{eff})	2,430
Compliance and outcome correlation	
Compliance rate (treated arm)	65%
Baseline–endline correlation (ρ)	0.8
Assumed effect sizes	
Expected ITT on mental health (H1, prior)	0.2–0.3 SD
Per-subscale ITT	0.10 SD
Per-subscale power (prior to RW correction)	
Analytical power	99.3%
DeclareDesign Monte Carlo power (5,000 sims)	99.3%
MDE at 95% power	0.080 SD
Family-wise power under Romano–Wolf (3,000 sims)	
H1 family—any subscale rejects	99.9%
H2 family—any subscale rejects	99.9%
H3a family—any subscale rejects	99.9%
H3b family—any subscale rejects	99.9%
H1 family—expected # rejections (of 7)	6.90
H2 family—expected # rejections (of 4)	3.95
H3a family—expected # rejections (of 3)	2.97
H3b family—expected # rejections (of 3)	2.97

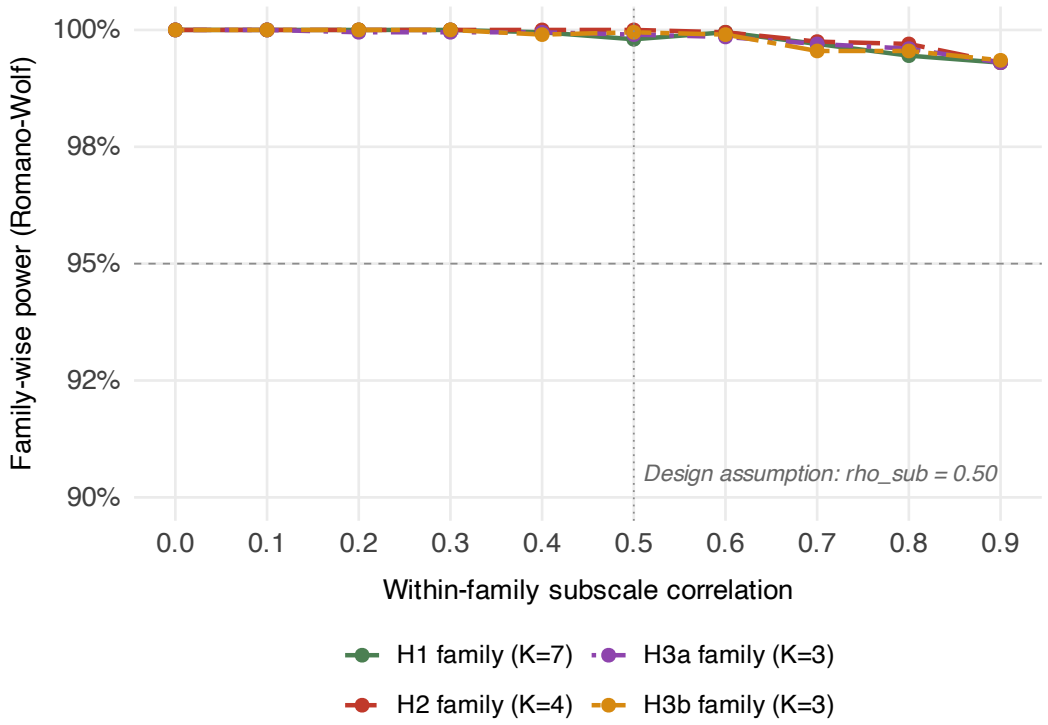
Note: Power computed in *DeclareDesign* (Blair et al., 2019) from full-DGP Monte Carlo simulations; see the preceding text for the data-generating process and estimator.

Figure D.1: Per-subscale power as a function of the baseline–endline correlation ρ.



Note: Per-subscale Monte Carlo power at the design ITT of 0.10 SD as ρ varies. The shaded region marks the range over which power falls below 95%. Computed via *DeclareDesign* on the same DGP described in the preceding text.

Figure D.2: Family-wise power as a function of the within-family subscale correlation ρ_{sub}.



Note: Family-wise power under the Romano–Wolf step-down at the design ITT of 0.10 SD, as the within-family between-subscale correlation ρ_{sub} varies. The vertical dotted line marks the design value $\rho_{\text{sub}} = 0.5$; the horizontal dashed line marks 95% power. Computed on the same DGP described in the preceding text.

E App installation and premium activation walkthrough, onboarding survey (E1)

Note. This is a preliminary version of the walkthrough: the exact in-app screens for the premium-activation step are not yet finalized, so the corresponding instructions are a placeholder that will be updated before fielding; the remaining steps are final.

Download Instructions

As part of this study, you'll be using a science-based well-being app called Flourish. Think of Flourish as your personal well-being buddy for enhancing both your personal and social well-being, backed by thousands of studies in psychology and behavioral science.

What you will do right now

1. Install the app on your smartphone and create an account.
2. During onboarding, select your organization and enter the organization code. This activates your free premium access automatically.
3. Enter your Flourish Buddy ID so we can confirm your participation.

To get started, please select which type of device you use:

o iPhone

o Android

LOGIC: If they select iPhone:

🌱 **Welcome to getting started with Flourish on iOS!** 🌱

We'll do this together, one step at a time.

First, please download the Flourish app onto your phone using the QR code below:



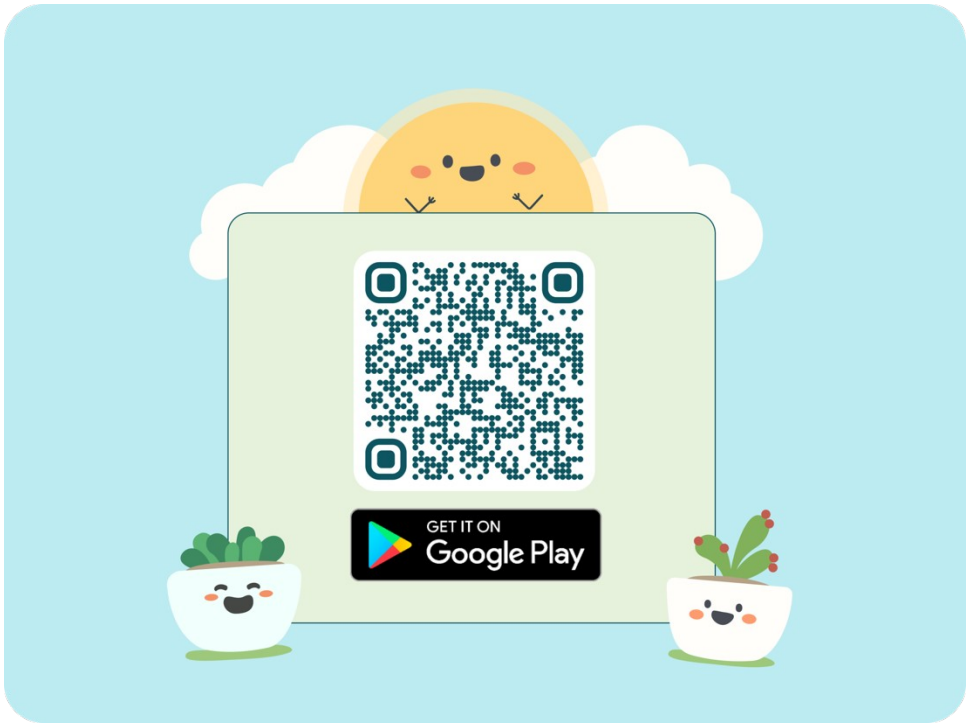
When the app has finished installing, open it and click "Next page >".

LOGIC: If they select Android:

🌱 **Welcome to getting started with Flourish on Android!** 🌱

We'll do this together, one step at a time.

First, please download the Flourish app onto your phone using the QR code below:



When the app has finished installing, open it and click "Next page >".

During the onboarding flow of the Flourish app, you will be asked how you are using Flourish.

Please select tap "I have an organization code" when prompted.

You will then be asked to enter your organization code. Please enter the following code exactly as shown:

${e://Field/OrgCode}

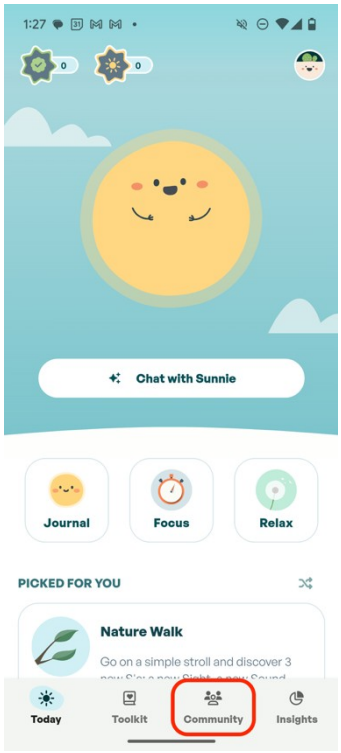
Once the code is accepted, your premium access (Flourish Plus) will be activated automatically for the three-month study period.

If a subscription or paywall screen appears at any point, you can close it using the button in the upper-right corner.

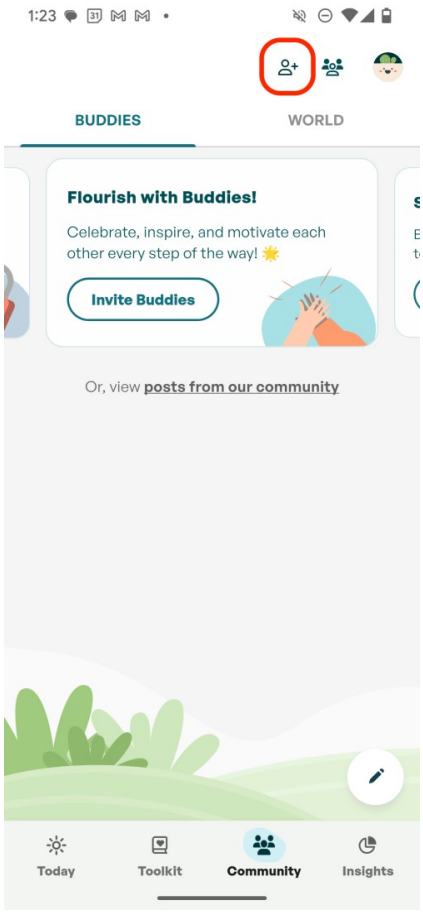
Continue through the rest of onboarding, then click "Next page >".

Next, please retrieve and enter your **6-character buddy code** from the app.

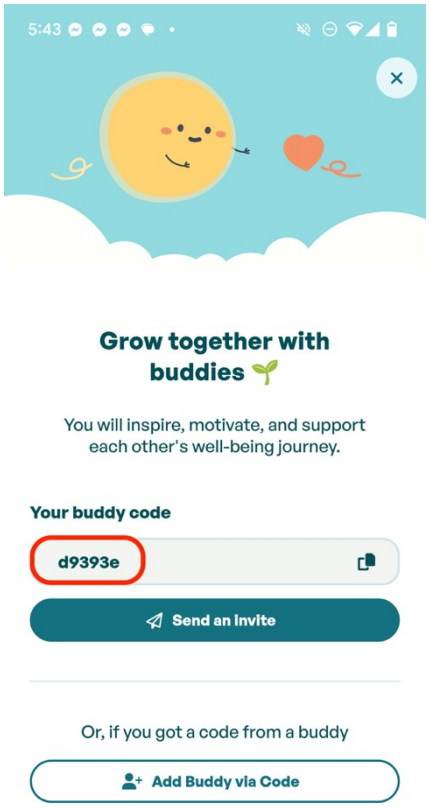
How to find your buddy code:
Open the app and navigate to the **Community** tab in the bottom menu bar:



Next, tap the icon on the upper right corner of your screen:



Finally, in the middle of the screen, you'll see your buddy code displayed -- that's the one to enter below!:



_____

📝 Once you've found it, please type it here:

_____

As a final step, please confirm each of the following:

- The Flourish app is installed on my phone and I have created an account.
- I entered the organization code and my premium access is active.
- I entered my Flourish Buddy ID.