Primary micro-randomized trial design and procedures

To evaluate the efficacy of the EMI delivered via a mobile app, a six-week micro-randomized trial (MRT) design was employed in the original study. This design was suitable for assessing the proximal effects of time-varying interventions and identifying how these effects change over time. In that trial, the decision point was defined as the moment immediately following the completion of the participant's daily evening Ecological Momentary Assessment (EMA). At each decision point, participants were randomized with a probability of 0.5 to receive either the 'Symptom-Tailored Intervention' (Intervention condition) or 'Active Control' condition. This longitudinal randomization enabled the estimation of the causal effect of providing symptom-tailored content on the subsequent psychological state.

The original study targeted South Korean undergraduate and graduate students aged 19 or older who were experiencing mild to moderate emotional difficulties. To ensure the intervention's focus on early prevention rather than clinical treatment, participants were required to meet the following inclusion criteria based on the PHQ-9 (Patient Health Questionnaire-9) and GAD-7 (Generalized Anxiety Disorder-7):

- Mild to Moderate Symptoms: A score of 5-14 on either the PHQ-9 or the GAD-7.
- Exclusion of Severe Symptoms: Participants scoring in the severe range (≥ 15) on both scales were excluded to ensure safety and intervention appropriateness.

The primary trial focused on undergraduate and graduate students who did not require clinical treatment but needed psychological support and were expected to receive practical help through intervention, aimed to examine the practical utility of preventive EMI.

Recruitment for the original trial was conducted from September to November 2024 through online communities at 31 South Korean universities. Advertisements were posted on internal university portals (e.g., Korea University's community) and 'Everytime', the most widely used campus community platform in South Korea. The primary experiment began in October and concluded in December. Each participant engaged for six consecutive weeks. Interested individuals underwent an initial online screening, and those who met the criteria were contacted via text, phone, or email to proceed. A dedicated mobile messenger was established for each participant to deliver ongoing study guidance, deliver interim surveys, and address technical inquiries or app errors throughout the trial. The original study was conducted remotely from recruitment to conclusion.

In accordance with the original MRT design, the decision point occurred immediately after participants completed the EMA response. At this point, participants were randomly assigned with a 50% probability to either the intervention condition or the control condition. The decision to deliver the intervention content at each decision point was performed independently by a computer-generated random sequence.

- Control Condition (50% probability): Participants responded to three items measuring state mindfulness, and the activity ended. This served as an active control condition, with no additional content provided.
- Intervention Condition (50% probability): Participants responded to the same three state mindfulness items as in the control condition. They also received tailored psychological education and content designed to help improve the symptom (depression, anxiety, stress, or sleep) identified as most vulnerable based on their EMA response for that day. The tailored content was provided by randomly selecting one of the contents targeting each symptom.

Due to the nature of the digital intervention, blinding of the participants was not feasible, as they could clearly perceive the delivery of the content. However, due to the MRT design, participants were likely to perceive the entire study period as a single continuous intervention experience rather than distinguishing between experimental and control states. This design naturally maintained a form of blinding, as users could not predict whether a lack of notification was due to the randomization result or the lack of a triggering context.

Primary micro-randomized trial sample size calculation

The required sample size for the original trial was determined using the MRT sample size calculator [1]. At the design stage of that experiment, assuming an accurate effect size was challenging due to the scarcity of comparable MRT studies. The investigators referred to a prior study on ACT interventions for first-generation college students, which assumed an effect size of 0.1 and confirmed significant effects on depression and stress [2].

In the primary study, a smaller effect size of 0.06 was assumed for the calculation, as a higher frequency of intervention repetitions was expected due to the limited number of content modules. Regarding availability, the EMA response rate from a 2023 pilot study using the same platform was 89%. Consequently, a more conservative availability rate of 80% was assumed for the MRT, which required an immediate EMI following the EMA. Based on these parameters, it was determined that a sample size of 240 participants would provide 80% power to detect a linear effect of 0.06 at a .05 significance level.

Intervention contents and decision rules

All provided content were adapted from materials originally developed by the Department of Psychiatry at Yongin Severance Hospital. They were designed to alleviate psychological distress, including depression, anxiety, and stress. The full list of intervention contents is presented in Table S1.

No.	Contents title	MRT decision rule (Target symptoms)	Rule-based, RL rule (action categories)
1	3-Minute Breathing Space	Depression, Anxiety, Stress, Sleep	Mindfulness
2	Mindfulness of Breath and Body	Depression, Anxiety, Stress	Mindfulness
3	Grounding	Anxiety, Stress	Mindfulness
4	Leaves on a Stream	Depression, Anxiety, Stress	Mindfulness
5	Raisin Meditation	Depression, Anxiety, Stress	Mindfulness
6	Mindful Eating Meditation	Depression, Anxiety, Stress	Mindfulness
7	Mindful Stretching (Mindful Yoga)	Depression, Anxiety, Stress, Sleep	Mindfulness
8	Sleep Hygiene Education	Depression, Stress, Sleep	CBT
9	Behavioral Activation	Depression, Stress	CBT
10	Thought Record	Depression	CBT
11	Relaxation Training (Abdominal Breathing)	Anxiety, Stress, Sleep	CBT
12	Resource Development and Installation	Depression, Anxiety, Stress	Positive Psychology
13	Self-Compassion Writing	Depression, Stress	Positive Psychology
14	Loving-Kindness Meditation	Depression, Stress	Positive Psychology
15	Imagery Training for Performance	Anxiety, Stress	Positive Psychology
16	Safe/Calm Place	Anxiety, Stress	Positive Psychology

Table S1: List of intervention contents

In the original MRT, when the randomization assigned a participant to the intervention condition, the specific content delivered was determined based on a pre-defined decision rule. The system identified the most severe symptom among depression, anxiety, stress, and sleep, as measured by the EMA immediately preceding the decision point. One intervention module was then randomly selected from a pool of content specifically targeting that primary symptom. This targeting logic was established by the clinical psychologists at Yongin Severance Hospital, who developed the materials. While all contents were expected to provide general benefits for depression, anxiety, and stress, this rule was implemented to maximize the proximal effect by addressing the user's most acute psychological vulnerability at that specific moment.

To evaluate the performance of the Reinforcement Learning (RL) model, a rule-based decision strategy was employed as a baseline. This strategy utilized the same action categories as the RL model (Mindfulness, Cognitive Behavioral Therapy [CBT], and Positive Psychology) but followed a fixed logic. Unlike the original MRT design, which randomized interventions at every decision point, the rule-based strategy provided content only when the participant's perceived need was high, as indicated by their real-time EMA scores. Specifically, the intervention was triggered only if at least one of the EMA scores for depression, anxiety, or stress was 5 or higher (Score ≥ 5). Once this threshold was met, one of the three intervention options was randomly selected and delivered. This rule-based approach served to compare the RL model's adaptive learning capabilities against a conventional, symptom-triggered intervention logic.

MDP formulation

We formulated the intervention decision problem as a finite-horizon Markov decision process. At participant i and day t, the observed state is denoted by $s_{i,t} \in R^d$, the action by $a_{i,t} \in \mathcal{A}$, and the reward by $r_{i,t} \in R$. Each transition is written as

$$(s_{i,t}, a_{i,t}, r_{i,t}, s_{i,t+1}). \tag{S1}$$

The action space consisted of four options,

$$\mathcal{A} = \{0, 1, 2, 3\}, \tag{S2}$$

where 0 denotes no content, and 1, 2, and 3 denote mindfulness, CBT, and positive psychology, respectively.

The reward was defined from next-day symptom improvement. Let $D_{i,t}$, $A_{i,t}$, and $S_{i,t}$ denote the depression, anxiety, and stress scores on day t. We defined

$$r_{i,t} = -\frac{1}{3}\left[(D_{i,t+1} - D_{i,t}) + (A_{i,t+1} - A_{i,t}) + (S_{i,t+1} - S_{i,t})\right]. \tag{S3}$$

Under this definition, a larger reward indicates greater next-day improvement in symptoms.

For a policy $\pi(a \mid s)$, the action-value function is

$$Q^{\pi}(s,a) = E_{\pi}\left[\sum_{k=0}^{T-t} \gamma^{k} r_{t+k} \,\middle|\, s_t = s,\ a_t = a\right], \tag{S4}$$

where $\gamma \in (0,1]$ is the discount factor. The corresponding state-value function is

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) Q^{\pi}(s,a). \tag{S5}$$

The optimal policy is defined as

$$\pi^{*}(s) = \arg\max_{a \in \mathcal{A}} Q^{*}(s,a). \tag{S6}$$

Offline RL objective

We learned the policy from logged trajectories only. Let the offline dataset be

$$\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{N}. \tag{S7}$$

The Q-network was trained by minimizing the temporal-difference regression loss

$$\mathcal{L}_{\mathrm{TD}}(\theta) = E_{(s,a,r,s') \sim \mathcal{D}}\left[\left(Q_{\theta}(s,a) - y(r,s')\right)^{2}\right], \tag{S8}$$

where the target is

$$y(r,s') = r + \gamma \max_{a' \in \mathcal{A}} Q_{\bar{\theta}}(s',a'). \tag{S9}$$

Here, $Q_{\bar{\theta}}$ denotes the target network.

If conservative Q-learning was used, we additionally included a conservative regularization term,

$$\mathcal{L}_{\mathrm{CQL}}(\theta) = \mathcal{L}_{\mathrm{TD}}(\theta) + \alpha\left[E_{s \sim \mathcal{D}} \log \sum_{a \in \mathcal{A}} \exp(Q_{\theta}(s,a)) - E_{(s,a) \sim \mathcal{D}} Q_{\theta}(s,a)\right], \tag{S10}$$

where $\alpha > 0$ is the regularization coefficient. This term penalizes overly large Q-values for actions that are weakly supported by the logged data.

After training, the learned policy was extracted by

$$\pi_{\mathrm{RL}}(s) = \arg\max_{a \in \mathcal{A}} Q_{\theta}(s,a). \tag{S11}$$

Rule-based baseline policy

As an interpretable comparator, we defined a rule-based policy using the same-day EMA symptom scores. Let D_t, A_t, and S_t denote the depression, anxiety, and stress scores at a given decision point. The rule-based policy was defined as

$$\pi_{\mathrm{rule}}(s_t) = \begin{cases} 0, & \text{if } \max(D_t, A_t, S_t) < 5, \\ \mathrm{Unif}\{1,2,3\}, & \text{if } \max(D_t, A_t, S_t) \geq 5. \end{cases} \tag{S12}$$

That is, if any of the three symptom scores was greater than or equal to 5, one intervention content was delivered by uniformly random selection from mindfulness, CBT, and positive psychology. Otherwise, no content was delivered.

Fitted Q-evaluation

To estimate the value of a target policy from logged data, we used fitted Q-evaluation (FQE). For a fixed target policy π, FQE iteratively learns an evaluation network Q_ϕ^π using Bellman targets induced by π.

At iteration k, the target for sample i is

$$y_i^{(k)} = r_i + \gamma \sum_{a' \in \mathcal{A}} \pi(a' \mid s_i')\, Q_{\phi^{(k-1)}}(s_i', a'). \tag{S13}$$

The network parameters are updated by minimizing

$$\phi^{(k)} = \arg\min_\phi \sum_{i=1}^{N} \left(Q_\phi(s_i, a_i) - y_i^{(k)} \right)^2. \tag{S14}$$

After convergence, the estimated policy value is computed from the initial states:

$$\hat{V}_{\mathrm{FQE}}^\pi = \frac{1}{N_0} \sum_{i=1}^{N_0} \sum_{a \in \mathcal{A}} \pi(a \mid s_i^{(0)}) Q_\phi(s_i^{(0)}, a), \tag{S15}$$

where $s_i^{(0)}$ denotes the initial state of trajectory i and N_0 is the number of evaluation trajectories.

Complementary off-policy estimators

Because a single FQE point estimate can be sensitive to function-approximation choices, we re-evaluated all three policies under a panel of complementary off-policy estimators. These comprised: (i) FQE with the main value-network architecture (MLP-128); (ii) FQE with an alternative, larger value-network architecture (MLP-256); (iii) a direct method (DM) based on a gradient-boosted next-day reward model fit on training participants; (iv) per-transition self-normalized importance sampling (PTWIS); and (v) weighted doubly robust estimation (WDR). For the importance-sampling–based estimators (PTWIS and WDR), we used both the design propensities from the MRT randomization and the empirically observed action propensities, yielding seven estimator variants in total. Estimators were averaged over policy-training seeds (42 and 43), and effective sample sizes were recorded for the importance-sampling estimators.

Across all seven estimator variants, the learned two-stage policy showed a higher estimated value than the behavior policy, with the learned-minus-behavior difference ranging from 0.16 (DM) to 0.42 (alternative FQE architecture; Fig. S1). The consistency of this ordering across estimators with differing bias–variance properties indicates that the advantage is not an artifact of any single estimation method, whereas the spread in magnitudes underscores that the absolute values should be read as estimates rather than precise quantities.

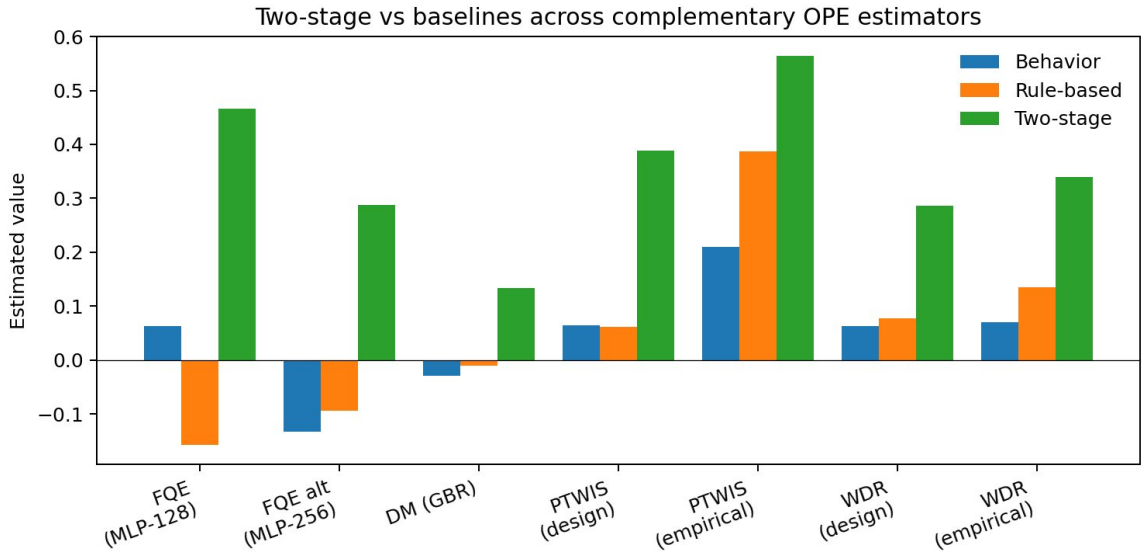

Figure S1: **Two-stage policy versus baselines across off-policy estimators.** Estimated value of the behavior, rule-based, and learned two-stage policies under seven estimator variants. The learned policy exceeded the behavior policy in all seven. FQE: fitted Q evaluation, DM: direct method, PTWIS: per-transition weighted importance sampling, WDR: weighted doubly robust.

Bootstrap procedure and results

Uncertainty in off-policy evaluation was quantified by nonparametric bootstrap resampling at the participant level. Participants were sampled with replacement, and all observations belonging to the same participant were kept together so that the within-subject temporal structure was preserved.

For bootstrap replicate $b \in \{1, \ldots, B\}$, let $\mathcal{D}^{(b)}$ denote the resampled dataset. The target policy value was re-estimated on each bootstrap sample:

$$\hat{V}_{\mathrm{FQE}}^{\pi,(b)} = \mathrm{FQE}(\mathcal{D}^{(b)}, \pi). \tag{S16}$$

The bootstrap mean and standard error were computed as

$$\bar{V}^{\pi} = \frac{1}{B} \sum_{b=1}^{B} \hat{V}_{\mathrm{FQE}}^{\pi,(b)}, \tag{S17}$$

$$\mathrm{SE}(\hat{V}^{\pi}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{V}_{\mathrm{FQE}}^{\pi,(b)} - \bar{V}^{\pi} \right)^2}. \tag{S18}$$

Percentile-based 95% confidence intervals were obtained from the empirical 2.5th and 97.5th percentiles of the bootstrap estimates.

We used $B = 1000$ replicates. The bootstrap distribution of the learned-minus-behavior policy-value difference lay entirely above zero (mean 0.646, 95% CI [0.599, 0.694]; Fig. S2), confirming that the ordering reported in the main text is stable under participant-level resampling.

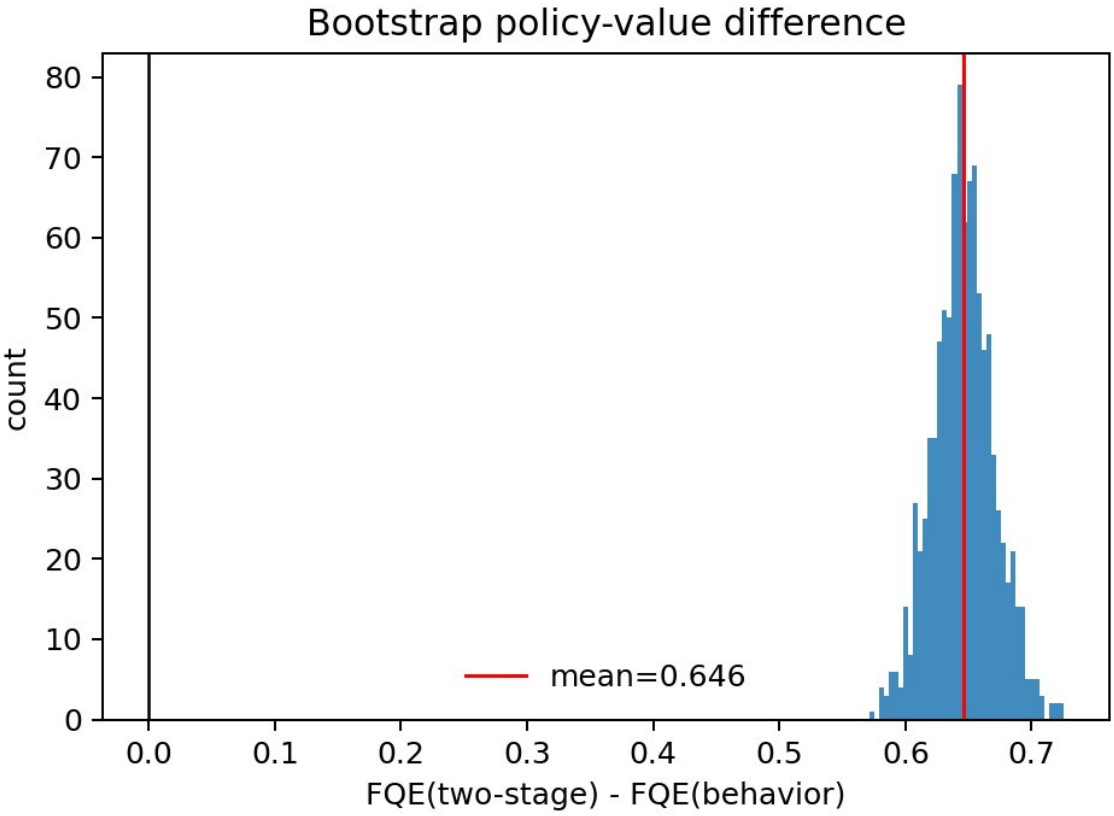

Figure S2: **Bootstrap distribution of the estimated policy-value difference.** Participant-level bootstrap (B=1000) distribution of the difference between the learned two-stage policy and the behavior policy. The distribution lies entirely above zero (mean 0.646, 95% CI [0.599, 0.694]).

Support and overlap diagnostics

A central concern in offline RL is whether the learned policy selects actions that are well represented in the logged data. Table S2 reports, for each action, the empirical (behavior) propensity and the propensity under the learned policy over all 7,980 decision points. The behavior data were not balanced across content types,

which is a direct consequence of the unequal number of items per category in the content pool and motivates the conservative learning approach. To check that the learned policy's shift toward positive-psychology content remained within the support of the data, we stratified states by quartiles of the first principal component (PC1, general psychological distress) and, for each action, computed the fraction of the learned policy's selections that fell into strata where the behavior policy chose that action less than 2% of the time ("thin-support" regions). For all four actions this fraction was 0.0, indicating that the learned policy did not place mass in regions where the corresponding action was essentially unobserved.

Action	**Behavior count**	**Behavior prop.**	**Learned count**	**Learned prop.**
Mindfulness	1934	0.242	338	0.042
CBT	986	0.124	690	0.086
Positive psych.	545	0.068	3157	0.396
No content	4515	0.566	3795	0.476

Table S2: **Action counts and propensities under the behavior and learned policies.** Propensities are computed over all 7,980 decision points. For every action, the fraction of learned-policy selections falling in thin-support PC1 strata (behavior propensity $< 2\%$) was 0.0. CBT: cognitive behavioral therapy.

One-stage versus two-stage factorization

We factorized the decision into a delivery gate and a content choice. This mirrors how a system actually operates: it first decides whether to interrupt the user at all, and only then selects content. The factorization also decouples a well-balanced binary decision (content vs. no content) from a content-type decision estimated using only the transitions on which content was actually delivered, where the unequal and limited per-category counts make conservatism essential.

To justify this choice empirically, we compared the two-stage policy against a single-stage four-action CQL model, sweeping the conservatism strength α over an identical grid and tracking both the no-content ratio and the FQE value (Fig. S3). The one-stage model had no setting of α that was simultaneously conservative and clinically usable. At small α, it over-delivered content—its no-content ratio fell to about 0.22, well below the behavior policy's 0.57—and its FQE value peaked but was inflated by overestimation of rarely delivered actions, precisely the failure mode conservatism is meant to prevent. As α increased, the one-stage model collapsed toward the in-data majority action: its no-content ratio rose past the behavior rate to nearly 0.98, and its estimated value fell below both the two-stage policy and the behavior policy, turning negative at the largest penalties. The two-stage model, by contrast, kept its no-content ratio within a controlled range (roughly 0.10 to 0.48, never exceeding the behavior rate) and its estimated value high and stable across most of the α range. This robustness follows from the factorization: Stage-1 learns the gating decision with class-rebalanced training, decoupling it from the data-level no-content imbalance, while Stage-2 is estimated only on transitions where content was actually delivered, so the no-content option toward which a single policy might collapse is structurally absent from the content-type decision.

Missing-data imputation

Missing data were imputed before policy learning. Let $x_t \in R^p$ denote the multivariate observation at day t, $m_t \in \{0,1\}^p$ the missingness mask, and $\delta_t \in R^p$ the time-gap feature indicating the elapsed time since the most recent observation for each variable. The imputation model takes the sequence

$$u_t = [x_t \odot m_t,\ m_t,\ \delta_t] \tag{S19}$$

as input and produces reconstructed values $\hat{x}_t$ for all time steps. The state-space model (Mamba) was applied causally (left-to-right, using only information up to time t, with no access to future observations) so that imputation could not introduce look-ahead leakage.

The imputation model was trained by randomly masking a subset of observed entries and minimizing reconstruction error only on the masked positions:

$$\mathcal{L}_{\text{imp}} = \frac{1}{|\Omega_{\text{mask}}|} \sum_{(t,j)\in\Omega_{\text{mask}}} |\hat{x}_{t,j} - x_{t,j}|\,, \tag{S20}$$

where Ω_{mask} denotes the set of artificially masked observed entries. After training, originally missing entries were replaced by the corresponding model predictions.

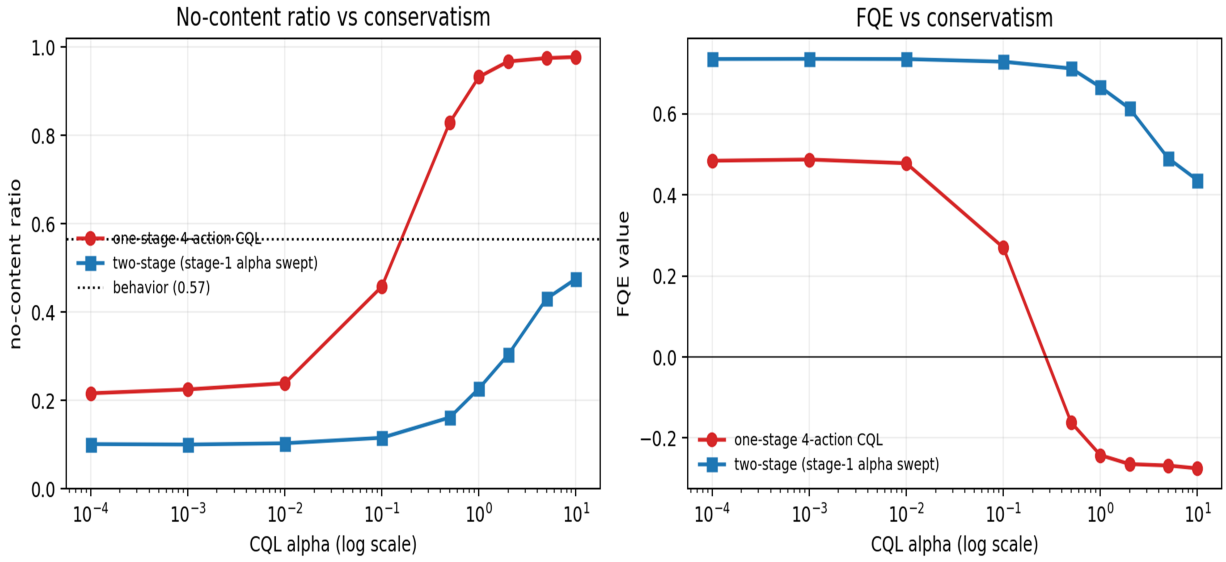

Figure S3: **One-stage versus two-stage policies across the conservatism strength.** No-content ratio (left) and FQE value (right) for the one-stage four-action CQL model (red) and the two-stage model with the Stage-1 penalty swept (blue), as a function of CQL α (log scale). The dotted line marks the behavior no-content ratio (0.57). FQE: fitted Q evaluation, CQL: conservative Q-learning.

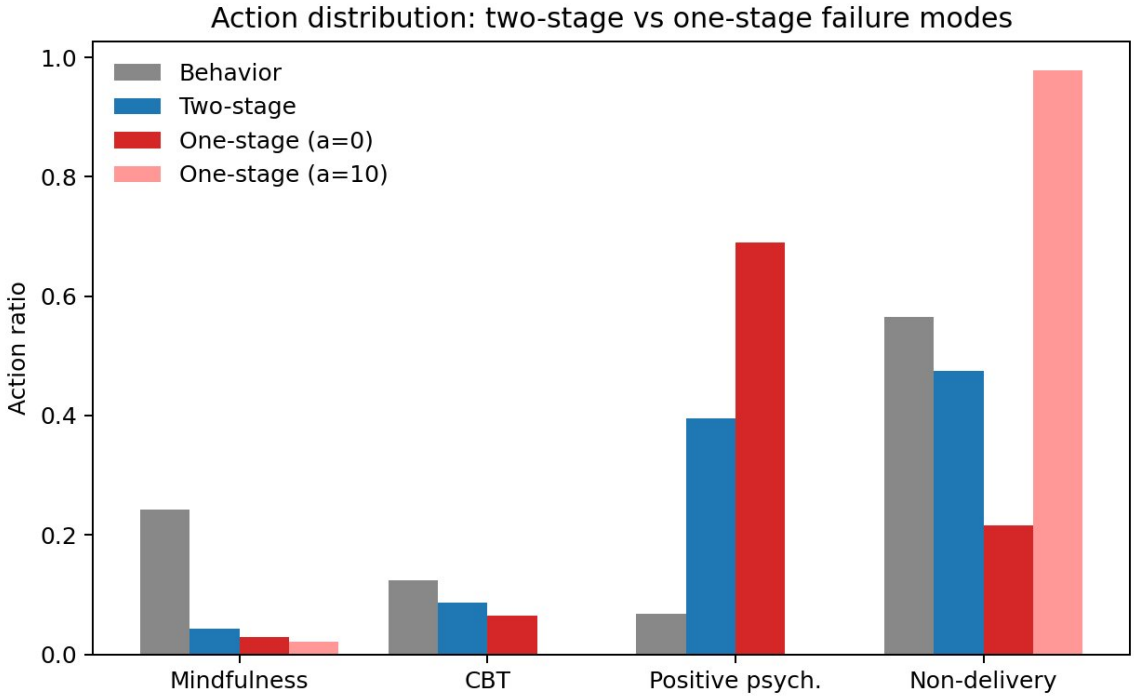

Figure S4: **Action distribution: two-stage policy versus one-stage failure modes.** Action ratios under the behavior policy, the two-stage policy, and the one-stage model at a low ($\alpha = 0$) and a high ($\alpha = 10$) conservatism setting. CBT: cognitive behavioral therapy, Positive psych.: positive psychology.

Sensitivity to imputation and missingness handling

Overall, 16.5% of rewards were affected by imputation, with 1,079 missing action logs and 2,827 rows containing at least one imputed state value. To assess the sensitivity of the main ordering to how missing data were handled, we re-estimated all three policies under six schemes that span the reviewer-requested variations (Fig. S5; Table S3). Across every scheme the learned two-stage policy retained a higher estimated value than the behavior policy, with the learned-minus-behavior gap ranging from +0.40 to +0.89, so the ordering does not depend on any single missing-data decision.

The schemes were as follows. The main analysis used the causal state-space (Mamba) imputation (two-stage

value 0.466). Restricting to complete observed transitions only, which drops rows containing any imputed or missing entry (5,146 of 7,980 transitions retained), preserved the ordering (0.493). Two simpler imputations were compared as a matched pair to probe data leakage: a global median imputation (0.623) and a strictly causal, train-only linear-regression imputation (0.614). The global median fills each column with its median computed over all participants and all time points, so it uses both held-out-participant information and future time points relative to a given decision (i.e., it is not leakage-safe); the linear-regression scheme instead fits on training participants only and predicts each missing value from same-time-point variables, with no access to future observations. That these two schemes yield nearly identical values (0.623 vs. 0.614) indicates that the leakage in the median scheme does not inflate the result—the difference between the policies is intrinsic rather than an artifact of leakage. Adding an explicit state missingness indicator left the ordering unchanged (0.626). Finally, to address the concern that coding missing actions as no-delivery may not be conservative, we excluded transitions whose action log was entirely missing (1,079 transitions removed, 6,901 retained); the gap actually widened (two-stage value 0.849), consistent with these missing-action transitions lying in easy "no-content" regions whose removal makes the learned policy's advantage more pronounced.

We did not perform an ablation comparing alternative imputation models for reconstruction accuracy; the causal state-space model was chosen because it uses only information available up to time t (no look-ahead), and the policy ordering was robust across the imputation schemes examined above. A more accurate imputation model may therefore exist, but the consistency of the ordering across schemes suggests the conclusions do not hinge on this choice.

Scheme	**Behavior**	**Rule-based**	**Two-stage**
Mamba (main)	+0.062	−0.158	0.466
Complete-only	−0.111	−0.253	0.493
Median†	+0.093	−0.165	0.623
Linear regression	+0.104	−0.125	0.614
Missingness indicator	+0.057	−0.137	0.626
Missing-action excluded	−0.044	−0.151	0.849

Table S3: **Sensitivity of the estimated policy values to imputation and missingness handling.** Estimated FQE values (averaged over policy-training seeds 42 and 43) for the behavior, rule-based, and learned two-stage policies under six schemes. In every scheme the two-stage policy exceeds the behavior policy. †The global median scheme is not leakage-safe (it uses all participants and all time points, including future ones); the causal linear-regression scheme is its leakage-safe counterpart, and the two give nearly identical two-stage values (0.623 vs. 0.614).

Sensitivity to the reward definition

Because the analytic dataset relied on a specific reward definition, we assessed the sensitivity of the main ordering to alternative reward constructions (Fig. S6; Table S4). The learned policy remained higher than the behavior policy in every case, but the magnitude of the gap varied considerably. Using next-day outcomes residualized on same-day symptoms—which removes a large part of any regression-to-the-mean effect—the learned value fell from 0.458 to 0.184, and single-dimension rewards yielded smaller gaps (depression 0.103, anxiety 0.077, stress 0.275). Conversely, longer-horizon rewards produced larger estimated values (two-day 0.540, three-day 0.648). As an additional check against simple mean reversion, the mean next-day change in the composite score was +0.136 following content delivery versus −0.041 under non-delivery, indicating a difference that exceeds the natural no-intervention trajectory in this sample. The magnitude—though not the direction—of the result therefore depends on the reward definition.

Reward definition	**Learned-policy value**
Composite raw DAS (main)	0.458
Residualized on same-day symptoms	0.184
Depression only	0.103
Anxiety only	0.077
Stress only	0.275
Two-day horizon	0.540
Three-day horizon	0.648

Table S4: **Sensitivity of the learned-policy value to the reward definition.** DAS: depression–anxiety–stress.

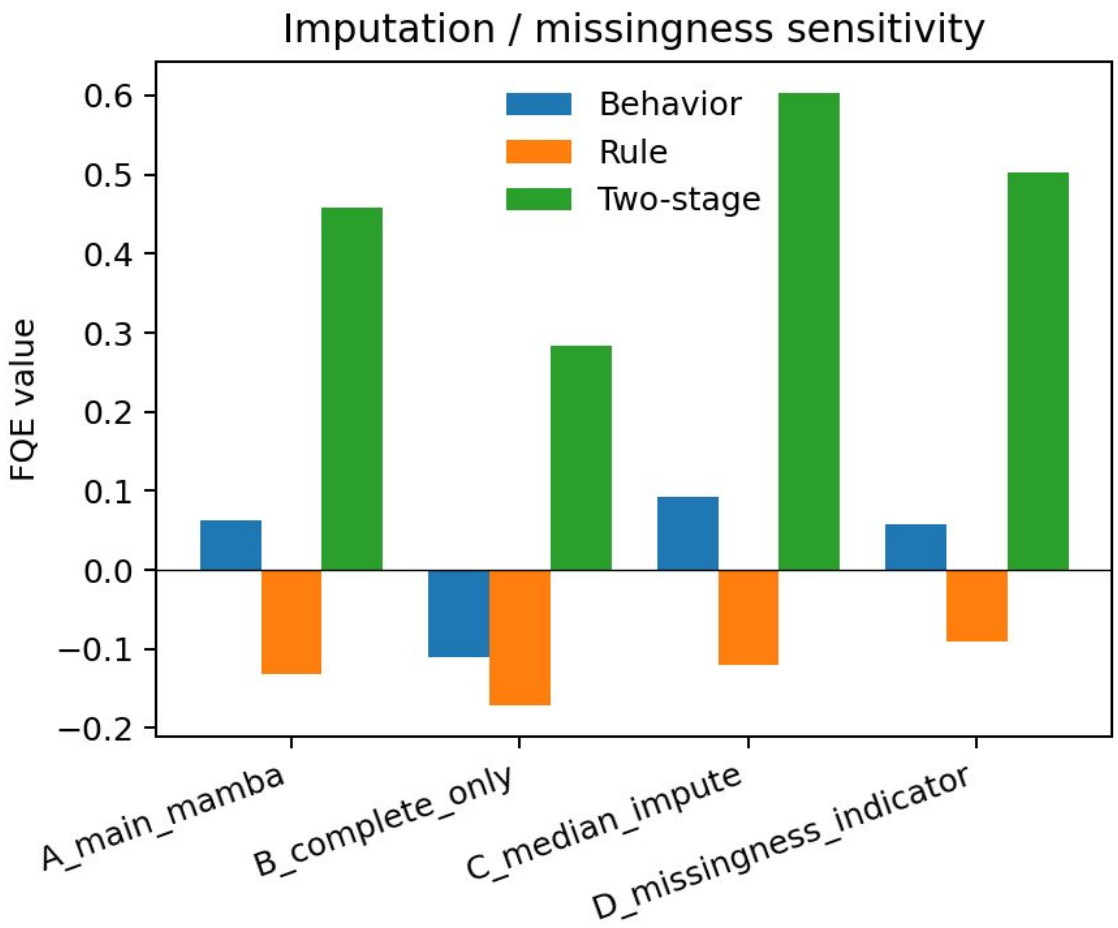

Figure S5: **Sensitivity to imputation and missingness handling.** Estimated FQE value of the behavior, rule-based, and two-stage policies under six schemes: causal state-space (Mamba) imputation, complete observed transitions only, global median imputation, causal linear-regression imputation, an added state missingness indicator, and exclusion of missing-action transitions. The learned two-stage policy exceeds the behavior policy in all six. FQE: fitted Q evaluation.

No evidence of temporal data leakage

To verify that the result was not driven by information leakage from the preprocessing pipeline, we re-ran the full analysis fitting every component (the imputation model, Z-score standardization, PCA, the reward model used by the direct and doubly robust estimators, and both CQL stages) on training participants only, and then evaluating FQE on a held-out set of test participants. The learned policy retained a higher estimated value than the comparators (0.387 vs. −0.096 for behavior and −0.055 for rule-based), indicating that the ordering does not depend on having seen test data during preprocessing.

Sensitivity to the participant inclusion threshold

The main analysis retained the 190 randomized participants who contributed at least 30 days of data and excluded 29 participants with shorter records. Because this threshold could, in principle, bias the learned policy toward more adherent users, we re-ran the entire pipeline including all 219 randomized participants under the identical protocol—the same policy-training seeds (42 and 43), hyperparameters, and train-only (leakage-safe) preprocessing as the main analysis. Including the 29 shorter-record participants extended the analytic dataset from 7,980 to 9,198 transitions. Under this inclusion criterion the learned two-stage policy retained a higher estimated value than both comparators (0.344 vs. −0.018 for behavior and −0.198 for rule-based); the learned-minus-behavior gap (0.362) was, if anything, slightly larger than in the main 190-participant analysis. The ordering is therefore robust to the 30-day inclusion threshold, indicating that the result is not an artifact of excluding lower-adherence participants. We nonetheless retain the adherence caveat in the main-text Limitations, since the excluded group remains too small to characterize low-engagement users in detail.

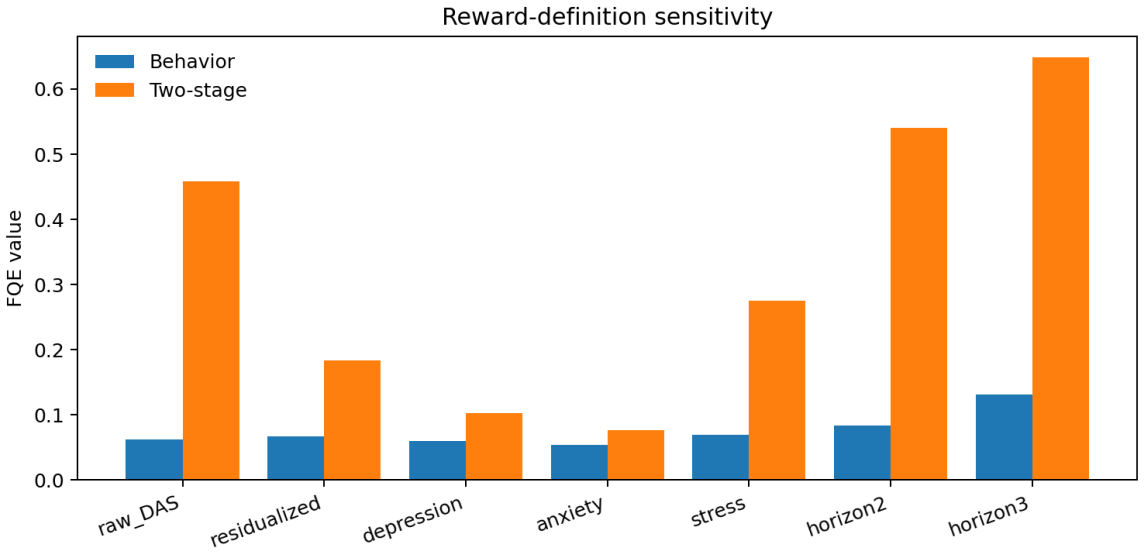

Figure S6: **Sensitivity to the reward definition.** Estimated FQE value of the behavior and two-stage policies under the composite raw DAS reward, a next-day outcome residualized on same-day symptoms, single-symptom rewards (depression, anxiety, stress), and longer-horizon rewards (two- and three-day). FQE: fitted Q evaluation, DAS: depression–anxiety–stress.

Nested model selection and training diagnostics

Hyperparameters were chosen by nested model selection so that the test set was never used for tuning. Participants were split 60/20/20 at the participant level; the state pipeline and both CQL stages were fit on the training participants, the two conservatism weights were selected on the validation participants by validation FQE over a grid ($\alpha_1 \in \{1, 5, 10, 20, 50\}$, $\alpha_2 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$), and the held-out test participants were used exactly once for the final evaluation. The selected configuration ($\alpha_1 = 10$, $\alpha_2 = 0.001$; validation FQE 0.671) yielded, on the untouched test set, an estimated value of 0.304 for the learned policy versus -0.196 for both baselines (Fig. S7). Across the validation grid, the learned-policy value was positive for nearly all configurations and degraded only at the most aggressive Stage-2 penalty ($\alpha_2 = 1$), consistent with excessive conservatism collapsing content delivery.

Training was stable and showed no signs of divergence (Fig. S8). The Stage-1 gating loss decreased monotonically (from 18.0 at epoch 1 toward 8.4), the Stage-2 content loss fell from 11.6 to 2.6 over 30 epochs, and the per-policy FQE temporal-difference loss decreased smoothly from approximately 12 to 2 within 40 epochs. The operating points used in the main analysis (Stage-1 1 epoch, Stage-2 10 epochs, FQE 2 epochs) were fixed by the validation procedure; longer training continued to reduce the temporal-difference loss but did not improve validation policy value, which is the expected behavior when the conservative penalty increasingly suppresses delivery.

Q-network and training details

The offline RL model used a feedforward Q-network that takes the state vector as input and outputs one Q-value for each action. The FQE model used the same architecture unless otherwise stated. The main hyperparameters included the number of hidden layers, hidden dimension, activation function, learning rate, batch size, discount factor, target-network update rule, and the conservative regularization coefficient when CQL was used.

PCA dimensionality

For visualization and interpretable subgroup analysis, the state representation was additionally projected onto a lower-dimensional PCA space. Let K denote the number of retained principal components. The cumulative explained variance was computed as

$$\mathrm{CEV}(K) = \frac{\sum_{j=1}^{K} \lambda_j}{\sum_{j=1}^{d} \lambda_j}, \tag{S21}$$

where λ_j is the eigenvalue of the jth principal component.

In the main analysis, $K = 6$ was used. The per-component and cumulative explained variance are reported in Table S5: the first six components together explained 46.8% of the total variance, and from PC7 onward each

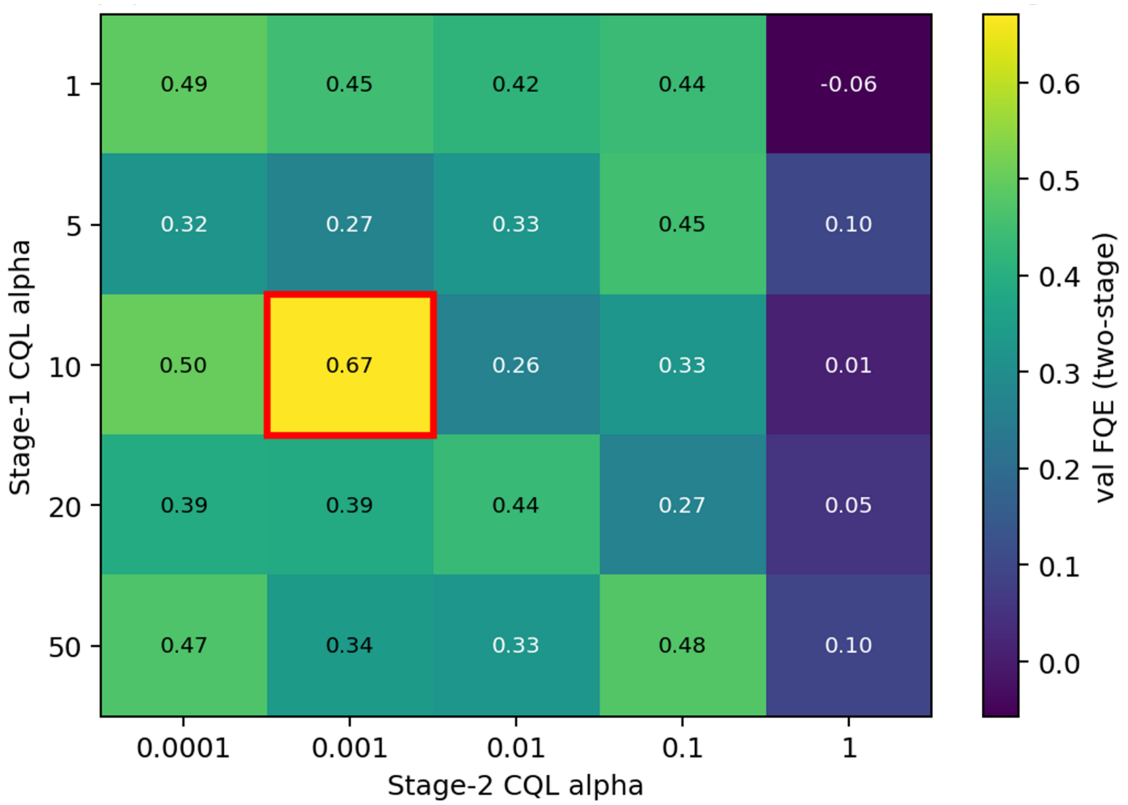

Figure S7: **Nested model selection over the conservatism weights.** Validation FQE of the learned two-stage policy across the Stage-1 (α_1) and Stage-2 (α_2) CQL penalty grid. FQE: fitted Q evaluation, CQL: conservative Q-learning.

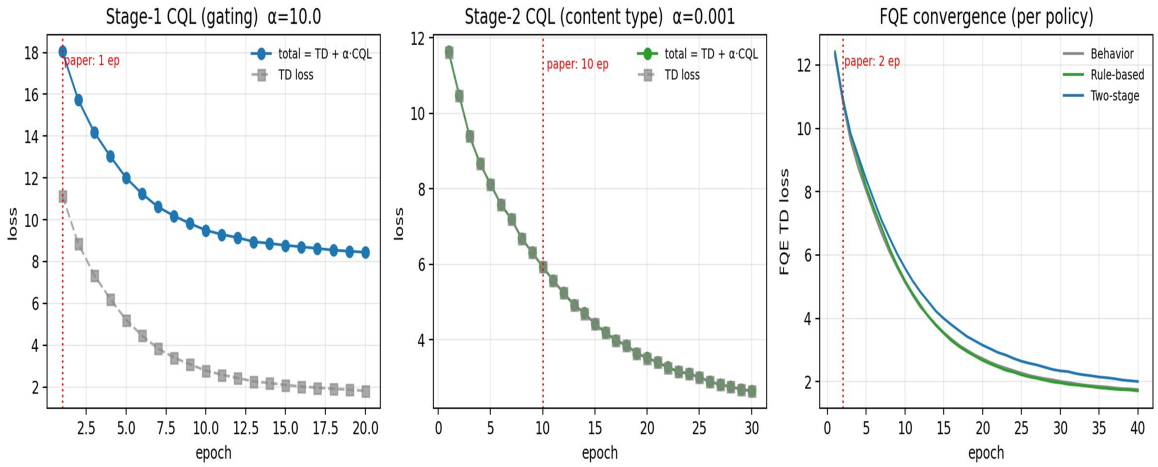

Figure S8: **Training and evaluation convergence diagnostics.** Loss curves for the Stage-1 CQL gating model, the Stage-2 CQL content model, and the per-policy FQE evaluation. CQL: conservative Q-learning, FQE: fitted Q evaluation, TD: temporal difference.

additional component contributed 4.3% or less with a gradual decline, placing the elbow near PC6.

Component	Explained variance	Cumulative
PC1	16.0%	16.0%
PC2	9.5%	25.5%
PC3	6.5%	32.0%
PC4	5.4%	37.4%
PC5	4.8%	42.2%
PC6	4.5%	46.8%
PC7	4.3%	51.0%
PC8	4.2%	55.2%
PC9	3.8%	58.9%
PC10	3.6%	62.5%

Table S5: **Explained variance of the principal components.** The first six components, retained in the main analysis, together explained 46.8% of the total variance.

PCA labels interpretation

The six components explained 46.8% of total variance (Table S2). Although this does not fully capture the variance in the dataset, we selected K = 6 to support convergent policy learning: the goal of dimensionality reduction here is not to reconstruct the original features faithfully but to provide a compact, low-noise state representation that supports stable Q-function approximation.

In this study, 29 state variables were compressed via principal component analysis (PCA) to derive six key axes. As PCA signs are mathematically arbitrary, all loadings should be interpreted relative to the assigned label direction. To assess the stability of each component across cross-validation folds, aligned cosine similarity was computed between PC axes estimated from training and test sets (Table S8). PC1 and PC2 demonstrated strong replication (aligned cosine = 0.97 and 0.90, respectively), while PC3 showed intermediate stability (0.66). PC4 through PC6 exhibited comparatively lower stability (0.27–0.43), and labels assigned to these components are therefore treated as provisional and interpreted with caution. It should be noted that PCA was employed primarily as a dimensionality reduction tool to enable interpretable summarization of the state space for the RL-based content delivery policy, rather than as a rigorous psychometric model. Researchers seeking to apply these components for strict inferential or clinical modeling purposes are advised to interpret them accordingly.

PC 1: General Psychological Distress

The first principal component is the axis of the total amount of subjective pain. Daily EMA variables, such as depression (0.322), stress (0.325), and anxiety (0.312), show large values in component loading, reflecting a subjective daily mood state. This is similar to the Depression, Anxiety and Stress Scale-21 (DASS-21) score (0.302, 0.288, and 0.302 for depression, anxiety, and stress sub-factors), which was measured every two weeks and updated. It can be seen that it is a typical psychological distress state in which not only is a simple negative emotion high, but also a lack of positive emotions, such as well-being (-0.322) and mood (-0.321).

PC 2: Psychosocial vulnerability and daily mood

The second component captures the contrast between the underlying vulnerability and the current adaptive state. Along with the lack of psychosocial protective factors such as self-esteem (-0.309) and perceived social support (-0.202), social emotional vulnerability such as the DASS-21 factors (depression 0.286, anxiety 0.286, stress 0.295), and social anxiety (0.329) are reflected as high loads. However, contrary to this, the well-being (0.298), mood (0.278), appetite (0.231), and sleep quality (0.210) of the EMA variables have a positive load, showing a contrasting spectrum between vulnerability and positive emotional states.

PC 3: Psychological resources and self-regulatory capacity

The third component represents the psychological resources as a basis for individuals to accept and execute intervening content. The higher the self-control (0.350), perceived social support (0.355), and self-esteem (0.268) are secured, the more this axis moves in the positive direction. In addition, it reflects the ability of smartphone dependence (-0.323) and academic procrastination behavior (-0.305) to perform tasks that require cognitive effort with negative loads.

Feature	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Sex	0.037	0.034	0.034	-0.245	0.020	0.197
Steps	-0.022	-0.039	-0.018	0.156	-0.189	-0.308
Circadian delay	0.001	0.018	-0.028	-0.029	0.092	-0.005
Circadian time	-0.036	0.012	0.008	-0.206	0.177	0.201
Appetite	-0.235	0.231	0.166	-0.051	0.154	-0.126
Depression	0.322	-0.150	0.153	-0.081	0.306	-0.103
Anxiety	0.312	-0.144	0.147	-0.079	0.315	-0.084
Stress	0.325	-0.211	0.089	-0.062	0.205	-0.022
Mood	-0.321	0.278	0.152	0.026	0.012	-0.061
Well-being	-0.322	0.293	0.175	0.019	0.033	-0.063
Sleep quality	-0.217	0.210	0.192	-0.112	0.318	-0.057
Sleep duration	-0.068	0.090	0.157	-0.115	0.408	0.049
Term	0.025	-0.020	0.028	0.109	0.020	-0.378
Total delivery	-0.061	-0.024	-0.174	-0.330	-0.147	0.370
Weekend	-0.045	0.080	0.068	-0.086	0.237	0.173
DASS-21 (D)	0.302	0.286	0.254	-0.010	-0.182	-0.015
DASS-21 (A)	0.288	0.286	0.299	-0.086	-0.181	-0.014
DASS-21 (S)	0.309	0.295	0.235	-0.021	-0.175	-0.039
SIAS	0.159	0.329	-0.177	-0.130	-0.028	0.052
GSDS	0.083	0.113	-0.042	0.272	-0.143	0.350
AUDIT	0.036	0.084	-0.035	0.522	0.225	-0.019
CDS	0.092	0.120	0.004	0.480	0.163	0.133
Smartphone	0.117	0.274	-0.323	-0.022	0.205	0.016
PSO	-0.123	-0.202	0.355	0.018	0.012	0.143
RSES	-0.147	-0.309	0.268	0.117	0.013	0.150
API	0.078	0.163	-0.305	-0.025	0.229	0.095
SCRS	0.001	-0.016	0.350	-0.026	-0.154	0.196
BPS	0.046	0.029	0.077	0.285	0.062	0.484

Table S6: **Factor loadings of all state variables on six principal components** Factor loadings of all 28 state variables on the six retained principal components (PC1–PC6). PHQ-9: Patient Health Questionnaire-9[3;4], GAD-7: General Anxiety Disorder-7[5;6], DASS-21: Depression, Anxiety and Stress Scale-21 Items[7;8], SIAS: Social Interaction Anxiety Scale[9;10], GSDS: General Sleep Disturbance Scale[11;12], AUDIT: Alcohol Use Disorders Identification Test[13;14], CDS: Cigarette Dependence Scale (Non-smokers were coded as 0)[15;16], Smartphone: Smartphone Addiction Proneness Scale (S-Scale)[17], PSO-8: Perceived Social Support through Others Scale-8[18], RSES: Rosenberg Self-Esteem Scale[19], API: Aitken Procrastination Inventory[20], SCRS: Self-Control Rating Scale[21;22], BPS: Boredom Proneness Scale[23;24].

PC 4: External reward seeking

The fourth principal component encompasses behavioral characteristics related to substance use. Alcohol dependence (0.522), smoking dependence (0.480), and boredom proneness (0.285) are strongly projected to reflect the sensitivity to external stimuli. This can represent the overall context in which it is difficult to be affected by intervention, which is also linked to a state in which the total number of intervention deliveries (-0.330) has a negative load and little content provision at the beginning of the intervention.

PC 5: Sleep-preservation-related psychological distress

In the fifth component, subjective depression (0.306) and anxiety (0.315) of the day were reported to be high, and at the same time, the quality of sleep (0.318) and sleep duration (0.408) indicators had a high load together. This means a state where subjective negative mood occurs, but the amount and quality of sleep are preserved. Unlike PC 1, which showed a typical level of psychological pain, daily distress with preserved sleep was identified as an independent dimension.

PC 6: Intervention fatigue

The sixth component reflects the late app usage time compared to the individual's biological rhythm (0.201) and boredom proneness (0.484). In addition, how many days have passed since the previous intervention (-0.378) has a negative load, and the accumulated total number of intervention provision (0.370) has a positive load. Late

Comp.	Interpretive label	Largest-magnitude loadings (sign)
PC1	Amount of general psychological distress	Stress (+.33), Depression (+.32), Well-being (−.32), Mood/Feeling (−.32), Anxiety (+.31), DASS-S (+.31), DASS-D (+.30), DASS-A (+.29)
PC2	Psychosocial vulnerability and positive daily mood	SIAS (+.33), RSES self-esteem (−.31), DASS-S (+.30), Well-being (+.29), DASS-A (+.29), DASS-D (+.29), Mood/Feeling (+.28), Smartphone (+.27)
PC3	Psychological resources and self-regulatory capacity	Social support PSO (+.35), Self-control SCRS (+.35), Smartphone (−.32), Procrastination API (−.30), DASS-A (+.30), RSES (+.27), DASS-D (+.25), DASS-S (+.24)
PC4	External reward seeking	AUDIT alcohol (+.52), CDS cigarette (+.48), Total delivery (−.33), Boredom BPS (+.28), GSDS sleep disturbance (+.27), Sex (−.24), Circadian time (−.21), Steps (+.16)
PC5	Sleep-preservation-related psychological distress	Sleep duration (+.41), Sleep quality (+.32), Anxiety (+.32), Depression (+.31), Weekend (+.24), Procrastination API (+.23), AUDIT (+.23), Stress (+.20)
PC6	High intervention fatigue	Boredom BPS (+.48), Term (−.38), Total delivery (+.37), GSDS sleep disturbance (+.35), Steps (−.31), Circadian time (+.20), Sex (+.20), SCRS (+.20)

Table S7: **Top loadings and labels of principal components** Top eight loadings by absolute value and assigned labels for each principal component. DASS: Depression Anxiety Stress Scale, SIAS: Social Interaction Anxiety Scale, RSES: Rosenberg Self-Esteem Scale, PSO: Perceived Social Support, SCRS: Self-Control Rating Scale, API: Aitken Procrastination Inventory, AUDIT: Alcohol Use Disorders Identification Test, CDS: Cigarette Dependence Scale, GSDS: General Sleep Disturbance Scale, BPS: Boredom Proneness Scale.

Component	Aligned Cosine
PC1	0.97
PC2	0.90
PC3	0.66
PC4	0.42
PC5	0.43
PC6	0.27

Table S8: **Stability of principal components across cross-validation folds** Aligned cosine similarity between PC axis derived from training and test sets for each of the six retained components.

use, boredom characteristic, the state of having just received intervention, and the state of having received a lot of intervention may have been combined to indicate a state of high intervention fatigue.

References

[1] Nicholas J Seewald, Ji Sun, and Peng Liao. Mrt-ss calculator: an r shiny application for sample size calculation in micro-randomized trials. *arXiv preprint arXiv:1609.00695*, 2016.

[2] Emily Brenny Kroska Thomas, Tijana Sagorac Gruichich, Jacob M Maronge, Sydney Hoel, Amanda Victory, Zachary N Stowe, and Amy Cochran. Mobile acceptance and commitment therapy with distressed first-generation college students: microrandomized trial. *JMIR Mental Health*, 10:e43065, 2023.

[3] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613, 2001.

[4] Seung-Jin Park, Hye-Ra Choi, Ji-Hye Choi, Kun-Woo Kim, and Jin-Pyo Hong. Reliability and validity of the korean version of the patient health questionnaire-9 (phq-9). *Anxiety and mood*, 6(2):119–124, 2010.

[5] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097, 2006.

[6] Jung-Kwang Ahn, Yeseul Kim, and Kee-Hong Choi. The psychometric properties and clinical utility of the korean version of gad-7 and gad-2. *Frontiers in Psychiatry*, 10:127, 2019.

[7] Peter F Lovibond and Sydney H Lovibond. The structure of negative emotional states: Comparison of the depression anxiety stress scales (dass) with the beck depression and anxiety inventories. *Behaviour research and therapy*, 33(3):335–343, 1995.

[8] Eun-Hyun Lee, Seung Hei Moon, Myung Sun Cho, Eun Suk Park, Soon Young Kim, Jin Sil Han, and Jung Hee Cheio. The 21-item and 12-item versions of the depression anxiety stress scales: psychometric evaluation in a korean population. *Asian nursing research*, 13(1):30–37, 2019.

[9] Richard P Mattick and J Christopher Clarke. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour research and therapy*, 36(4):455–470, 1998.

[10] Sojung Kim, Yoon Hyae Young, and Jung-Hye Kwon. Validation of the short form of the korean social interaction anxiety scale (k-sias) and the korean social phobia scale (k-sps). *Cognitive Behavior Therapy in Korea*, 13(3):511–535, 2013.

[11] Kathryn A Lee. Self-reported sleep disturbances in employed women. *Sleep*, 15(6):493–498, 1992.

[12] Heejung Choi, Sung-Jae Kim, Beomjong Kim, and Inja Kim. Korean versions of self-reported sleep questionnaires for research and practice on sleep disturbance. *Korean Journal of Rehabilitation Nursing*, pages 1–10, 2012.

[13] Thomas F Babor, John C Higgins-Biddle, John B Saunders, Maristela G Monteiro, World Health Organization, et al. Audit: the alcohol use disorders identification test: guidelines for use in primary health care. Technical report, World Health Organization, 2001.

[14] Byung Ook Lee, Choong Heon Lee, Pil Goo Lee, Moon Jong Choi, and Kee Namkoong. Development of korean version of alcohol use disorders identification test(audit-k)its reliability and validity. *Journal of Korean Academy of Addiction Psychiatry*, 4(2):83–92, 2000.

[15] Jean-Francois Etter, Jacques Le Houezec, and Thomas V Perneger. A self-administered questionnaire to measure dependence on cigarettes: the cigarette dependence scale. *Neuropsychopharmacology*, 28(2):359–370, 2003.

[16] CY Park, JH Ahn, SY Bang, and SW Choi. The reliability and validity of the korean version of the cigarette dependence scale-12 (cds-12). *Journal of Korean Academy of Addiction Psychiatry*, 14(1):42–48, 2010.

[17] Dongil Kim, Chung Yeoju, Lee Juyoung, Myeungchan Kim, Yun-Hee LEE, Eunbi Kang, Chang-Min KEUM, and Jee-Eun Karin Nam. Development of smartphone addiction proneness scale for adults: Self-report. *Korea Journal of Counseling*, 13:629–644, 2012. doi: 10.15703/kjc.13.2.201204.629.

[18] K. Haram, K Jaewon, and K NA-Rae. Development of short perceived social support through others scale (pso-8): A rasch analysis. *The Journal of Human Understanding and Counseling*, 42(1):51–70, 2021.

[19] Ha-Na Bae, Sam-Wook Choi, Je-Chun Yu, Jong-Sun Lee, Kyeong-Sook Choi, et al. Reliability and validity of the korean version of the rosenberg self-esteem scale (k-rses) in adult. *Mood Emot*, 12(1):43–49, 2014.

[20] Margaret Eppinger Aitken. *A personality profile of the college student procrastinator*. University of Pittsburgh, 1982.

[21] Philip C Kendall and Lance E Wilcox. Self-control in children: development of a rating scale. *Journal of Consulting and Clinical psychology*, 47(6):1020, 1979.

[22] Won-young Song. Effects of self-efficacy and self-control on the addictive use of internet. *Unpublished master's thesis, Yonsei University, Seoul*, 1998.

[23] Richard Farmer and Norman D Sundberg. Boredom proneness–the development and correlates of a new scale. *Journal of personality assessment*, 50(1):4–17, 1986.

[24] KH Kim. The mediation effect of the ability to perceive emotion on boredom proneness and mental health. *Unpublished master's thesis, Catholic University*, 2012.