Human Error, Reliability, Resilience, and Performance, Vol. 213, 2026, 67–77
https://doi.org/10.54941/ahfe1007555

Designing Personalized Feedback Comments for Safety Management Based on Personality and Engagement: The PersonaTrace Scope Framework

Kaito Takahashi, Yu Shibuya, and Yusaku Okada

Graduate School of Science and Technology, Faculty of Science and Technology, Keio University, Japan

ABSTRACT

In high-risk work systems, safety is increasingly sustained not by procedural compliance alone but by operators' continuous interpretation of unfolding situations, attention allocation, and adaptive coordination under time pressure. Yet operationalizing personalized safety interventions remains challenging due to (i) poorly specified personalization dimensions, (ii) unstable message quality at scale, and (iii) limited deployability in real organizations. We propose PersonaTrace Scope (PTS), a two-axis design framework that specifies what to say, to whom, and at what cognitive depth. PTS independently controls (a) message framing and tone through seven practice-oriented personality factors and (b) intervention depth—defined as behavioral demand and expected cognitive load—through five engagement levels. This yields 175 predefined scopes with corresponding structured comment templates. To sustain minimum quality across this corpus without exhaustive expert review, we use a large language model (LLM) strictly as a quality assessor—not as a content generator—and introduce the Behavioral Trigger Index (BTI), a multidimensional gate that quantifies action-triggering properties relevant to human reliability (self-efficacy, adaptability, immediate executability, self-relevance, and organizational influence). Comments are iteratively refined until they satisfy a prespecified BTI threshold. We conducted a field evaluation with approximately 150 nurses, assessing perceived agreement, usefulness for patient safety activities, and intention to apply the content during work. Positive responses (Agree/Strongly agree) were 80%, 72%, and 74%, respectively, indicating that PTS-delivered comments are acceptable and may support safety-related reflection and behavioral intention. PTS offers an operational pathway to reconcile personalization with scalable quality assurance in high-risk domains.

Keywords: Personalization, Safety adaptability, Human reliability, Engagement, Feedback

INTRODUCTION

In contemporary high-risk industries, safety performance increasingly depends on frontline workers' capacity to maintain control in dynamic, uncertain conditions. When the work system deviates from nominal assumptions—through variability in workload, interruptions, patient or customer acuity, or resource constraints—operators must prioritize, monitor for weak signals,

Received February 5, 2026; Revised April 4, 2026; Accepted April 19, 2026; Available online July 20, 2026

and adjust actions while coordinating with others. These places sustained demands on attention management and cognitive workload, and make human reliability a function of judgment and adaptation rather than rule-following alone. As a result, safety is being reframed as an emergent capability achieved through adaptive performance.

We conceptualize this capability as safety adaptability: the capacity to maintain or recover safety under ambiguous or unexpected conditions by updating situation assessment, selecting compensatory actions, and coordinating with others. Although safety research has discussed related constructs such as self-efficacy, autonomy, and safety citizenship, practical and deployable principles for designing daily individualized micro-interventions that support adaptive performance remain insufficiently organized.

Digital technologies have accelerated personalization and just-in-time adaptive interventions across behavior-change domains. However, transferring these approaches to safety management faces three persistent challenges. First, personalization often lacks well-defined design axes and degenerates into superficial branching by attributes, without a principled link to attention, cognitive load, or acceptability. Second, quality assurance is difficult to scale: generic or vaguely worded advice can be ignored, while overly prescriptive messaging can increase mental load, trigger reactance, and undermine responsibility ownership. Third, operational scalability is constrained by the cost of authoring, updating, and auditing large numbers of tailored messages.

To address these challenges, we propose PersonaTrace Scope (PTS), an operational framework that explicitly designs "what to say, to whom, and how deeply." PTS separates personalization into two independent axes: (i) personality-related tendencies that shape how messages are framed and received (tone, perspective, and rhetorical stance), and (ii) engagement level that determines tolerable intervention depth and expected behavioral demand, thereby controlling cognitive and motivational burden. This separation supports flexible design—e.g., preserving the same safety-critical content while adapting tone, or increasing behavioral demands while maintaining a stable relational stance.

Using seven personality factors and five engagement levels, we designed 175 scoped comment patterns as structured templates that embed action-triggering elements (specificity, feasibility, and psychological acceptability). We further introduce a scalable quality assurance loop in which an LLM is used strictly as an evaluator—a quality gate—rather than a generator. We operationalize minimum quality using the Behavioral Trigger Index (BTI), and iteratively refine comments until they satisfy prespecified criteria. The contributions of this study are threefold: (1) a two-axis personalization design for safety micro-interventions (personality-based framing × engagement-based depth control), (2) a quantitative, scalable quality gate for ensuring minimum comment quality across a large template corpus, and (3) empirical evidence on acceptability and behavioral intent from a healthcare field evaluation. We next describe the PTS framework, scope allocation procedure, BTI-based quality assurance loop, and field results.

Designing PersonaTrace Scope (PTS): A Two-Axis Framework for Operationalized Personalized Safety Interventions

Design Requirements and Overall Framework

The interventions targeted in this study are not one-off training programs or temporary awareness campaigns; they are brief, recurring prompts delivered within day-to-day operations. Therefore, the design must be both theoretically grounded and operationally feasible under real constraints (time pressure, interruptions, and limited attention). Based on prior work and practical considerations, we prioritize three requirements. First, acceptability. Even well-intended safety guidance will be disregarded if it clashes with the recipient's values, decision style, or perceived autonomy. In safety-critical contexts, messaging must preserve psychological ownership and responsibility rather than merely prescribing correct behavior. This implies deliberate control over tone, framing, and perceived interpersonal stance. Second, actionability. Abstract cautionary statements are easy to produce but often fail to translate into behavior. Effective prompts should reduce cognitive load by specifying a small, executable action or decision cue that can be enacted under real constraints (e.g., when attention is divided and the task is time-bounded). Third, operational scalability. Fine-grained personalization increases costs for content creation, revision, and governance. A workable system needs a bounded design space with reproducible assignment rules and a quality assurance mechanism that does not depend on continuous manual expert review.PTS addresses these requirements by constraining personalization to meaningful, controllable design dimensions. The end-to-end operational flow (input → scoring → classification → scope allocation → comment assignment → output) is shown in Figure 1.

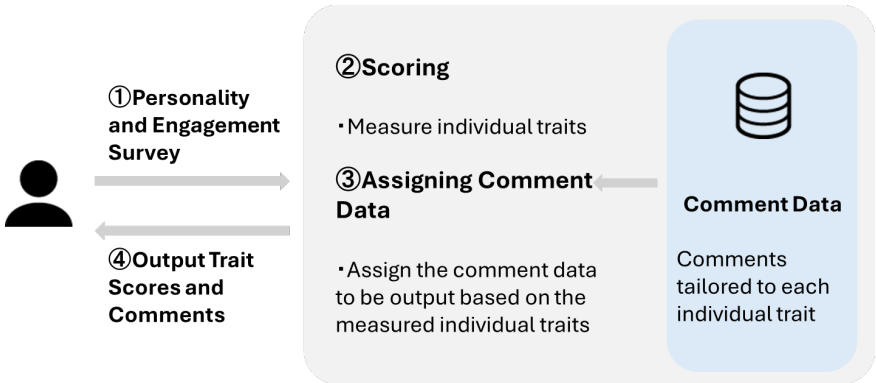

Figure 1: The operational flow of PTS.

Two-Axis Personalization and Template Design Across the Defined Space

The core idea of PTS is to represent personalization through two independent axes rather than a single undifferentiated "degree of tailoring." The first axis is personality-related tendencies that shape what types of framing are likely to be acceptable and persuasive. For example, reflective individuals may respond better to reasoning-oriented prompts that acknowledge uncertainty and encourage deliberate checks, whereas socially oriented individuals

may respond better to direct, collaborative language that frames safety as a shared endeavor. Importantly, in PTS, personality factors primarily govern expression—tone, perspective, and rhetorical strategy—rather than determining the safety content itself.The second axis is engagement level, defined as the individual's readiness to invest effort in improvement activities and to tolerate behavioral demand. Engagement determines the acceptable depth of intervention: high engagement allows prompts that invite proactive risk anticipation, reflective learning, and influence on others; low engagement calls for minimal, low-burden prompts that reduce friction and focus attention on one critical behavior. This axis explicitly manages expected cognitive load and the likelihood of reactance, aligning intervention depth with the recipient's motivational capacity and attention resources. Combining seven personality factors with five engagement levels yields 175 scopes. This granularity is not intended as maximal personalization; it is a deliberately bounded resolution that remains meaningful and manageable in operational settings. The purpose is not clinical diagnosis but an explainable, reproducible reference frame for selecting an intervention style and depth. The overall structure and creation flow of the comment corpus are shown in Figure 2.

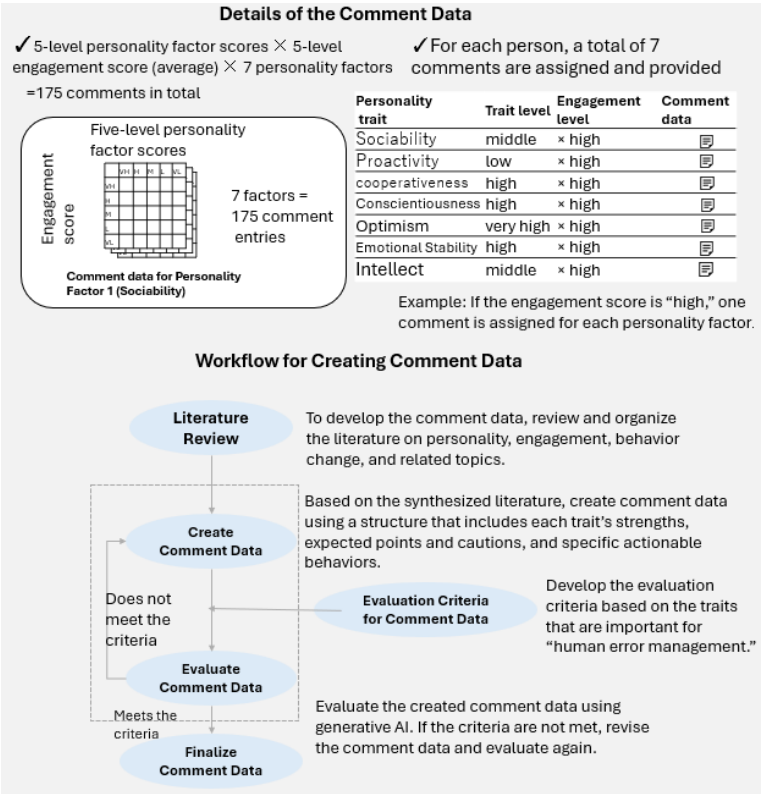

Figure 2: Details of the comment dataset and the creation process.

For each scope, we authored a corresponding safety comment using a fixed template structure rather than unconstrained generation. Each comment includes:

(1) a brief contextual acknowledgement or alert that orients attention;
(2) articulation of the safety-critical judgment/behavior;
(3) an immediately executable step or cognitive cue that reduces decision ambiguity; and
(4) when appropriate, a suggestion for proactive coordination or influence on others (e.g., speaking up, sharing concerns, or aligning expectations).

Personality primarily shapes the expression of (1) and (2), while engagement governs the demand and depth of (3) and (4). By fixing structure and controlling variable components, PTS aims to combine personalization with consistent minimum quality and predictable cognitive burden.

PTS should thus be understood not as a text-generation method but as an operational design framework for safety micro-interventions. Compared with one-size-fits-all messaging, PTS provides a principled way to account for individual differences. Compared with unconstrained optimization approaches, it maintains feasibility by bounding the design space and enabling systematic quality governance. It reframes "What should we say to whom?" into "Which predefined scope should we select?"—making intervention selection a reproducible design decision rather than an ad hoc judgment.

Profiling Individuals and Allocating Personalized Intervention Scopes

This section describes how PTS is applied in practice: how we measure personality tendencies and engagement and map individuals to one of the 175 scopes (see Figure 1). The profiling is not a clinical assessment; it is a pragmatic procedure intended to support consistent scope allocation.

Profiling Instrument and Measures

We administered a 113-item questionnaire with two parts: (a) items assessing personality-related tendencies relevant to safety performance (behavioral tendencies, decision style, and interpersonal involvement), and (b) items assessing engagement related to work and safety improvement (initiative, ownership, and improvement orientation). Items were rated on a 5-point Likert scale.

Personality was operationalized using seven practice-oriented trait factors that capture differences commonly observed in safety-relevant behavior. Scores were computed as weighted means of associated items, including reverse-keyed items to mitigate response bias. Engagement was treated as a composite indicator of involvement in safety and improvement activities, representing how much intervention depth and behavioral demand the individual is likely to accept.

Scoring, Aggregation, and Discretization

Responses were converted to numerical scores: +2 (strongly agree), +1 (agree), 0 (neutral), –1 (disagree), and –2 (strongly disagree). Reverse-keyed items were scored with inverted signs. Item weights (0.0–1.0), determined through expert review, were applied to compute weighted factor scores:

$$Score_f = \frac{\sum_{i=1}^{nf} \omega_i s_i}{\sum_{i=1}^{nf} \omega_i}$$

where s_i is the scored response for item i, ω_i is the corresponding weight, and n_f is the number of items in factor f. Scores were normalized within factors to support comparability across individuals. Missing values were imputed using the mean of valid responses within the same factor; if missingness exceeded a predefined threshold, that factor score was treated as invalid, and no scope was assigned for that factor.

Normalized scores were discretized into five levels for each personality factor and for engagement. Thresholds were set based on observed distributions and adjusted to avoid extreme class imbalance. Engagement levels were interpreted as design constraints: lower levels correspond to minimal, low-burden prompts; higher levels permit deeper prompts that require more reflection, initiative, and proactive coordination.

Scope Allocation

For each individual, we identified the most salient personality factor among the seven (i.e., the factor most pronounced relative to the individual's profile). We then combined (i) that factor's five-level category and (ii) the individual's five-level engagement category to select the corresponding scope from the 175-scope matrix. This yields one representative scope per person, enabling a straightforward and reproducible mapping between profiling outputs and intervention delivery. The next section details how comments associated with each scope are quality-assured using an LLM-based assessment and the Behavioral Trigger Index (BTI).

Ensuring Intervention Quality at Scale: LLM-Based Evaluation and the Behavioral Trigger Index

Automated Quality Gate and BTI Overview

PTS entails a large pre-authored corpus (175 scoped comment patterns). Ensuring a consistent minimum quality level across this set is difficult to sustain through manual expert review alone. In safety contexts, low-quality messaging is not merely ineffective; it can increase cognitive load, dilute attention to critical cues, provoke reactance, and shift responsibility in undesirable ways. We therefore implement an automated quality gate to screen and refine comments prior to deployment.

Crucially, the LLM is used strictly as an evaluator—not as the author of comments. We quantify comment quality using the Behavioral Trigger Index (BTI), a composite measure intended to capture whether a comment has the

properties required to trigger safety-relevant behavior under operational constraints. BTI comprises five dimensions (D1–D5): self-efficacy, situational adaptability, immediate executability, self-relevance, and organizational influence. Figure 3 summarizes the definition and full-score criteria for each dimension. Each dimension is rated on a 0–100 scale via a fixed evaluation prompt, and BTI is computed as a weighted aggregation that emphasizes action-enabling content. We apply BTI as a gatekeeping threshold (BTI ≥ 70) to enforce a minimum standard of operational validity and cognitive fit, while keeping evaluation settings fixed (prompt format and generation parameters) to improve reproducibility.

Evaluation Metric	Weight	Capability	Definition & Evaluation Focus	Criteria for Full Score
D_1: Self–efficacy (self–efficacy)	–	An attitude of engaging proactively without denying one's traits	Does it include affirmations tailored to the person's traits? Does it accept the recipient's personality traits and current state without negation, using an accepting tone?	The text can affirm the recipient's personality traits, and the recipient can feel strong empathy/identification.
D_2: Situational adaptability (autonomous practice capability)	0.4	Noticing change and flexibly judging/adjusting	Are "decision criteria" and/or "adaptation strategies" for dealing with changes presented? Does presenting judgment criteria increase the range of effective responses the recipient can draw on?	Provides a concrete strategy for how to adjust one's traits to the situation.
D_3: Immediate executability (autonomous practice capability)	0.4	The ability to initiate specific actions immediately	Are minimal, immediately actionable behaviors presented that the recipient can start right away? Does suggesting a minimal action lower the barrier to action?	Low-burden, concrete actions/movements are presented.
D_4: Self-relevance (autonomous practice capability)	–	Processing information as personally relevant	Is it written in a tone consistent with the recipient's personality? Does that help reduce psychological resistance and increase comprehension?	The wording matches the recipient's personality.
D_5: Organizational influence (organizational influence capability)	0.2	Extending individual actions to the team-wide culture	Does it go beyond individual action and connect to sharing within the team and to future career formation? Does it promote positive spillover effects on others?	It describes how the individual's improvement action contributes to transforming organizational culture and to one's own growth.

Figure 3: Definition and full-score criteria for each dimension.

Iterative Refinement and Reliability Considerations

BTI is applied within an iterative refinement loop: comments are minimally revised and re-evaluated until they meet the threshold. Revisions are constrained to the predefined template structure to preserve the intended intervention design and to avoid overfitting to the BTI score. To assess robustness, we compared BTI outputs across two LLMs and observed broadly similar scoring tendencies, suggesting that evaluation is not dominated by idiosyncrasies of a single model. BTI is not intended to replace outcome evaluation. Rather, it functions as a scalable first-line quality assurance mechanism that enforces minimum standards before workplace exposure. The effectiveness of the delivered comments is assessed via the subsequent field evaluation. Figure 4 summarizes BTI score trajectories across refinement iterations, showing that the corpus consistently meets the gatekeeping threshold (BTI ≥ 70) after the quality assurance loop.

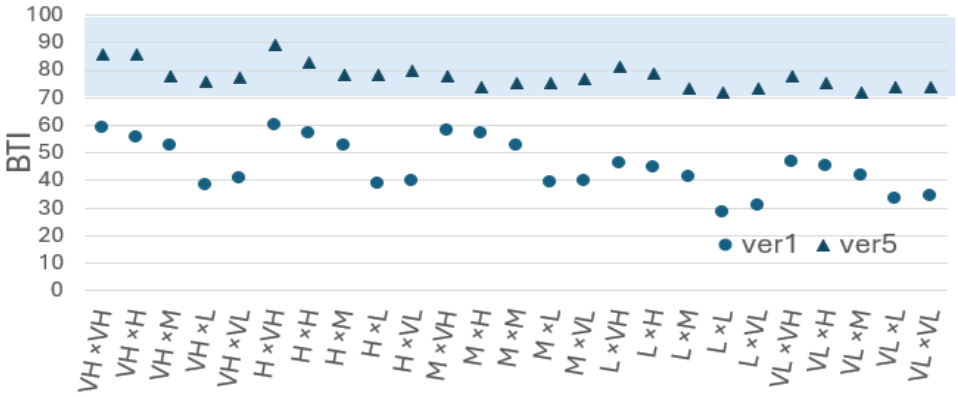

Figure 4: BTI score improvement.

Field Evaluation: Acceptance and Behavioral Intent in a Clinical Safety Context

Study Context and Procedure

To examine feasibility in an applied setting, we conducted a field evaluation with approximately 150 nurses in a domestic hospital. Nursing was selected because safety-critical work is embedded in routine practice and is characterized by high interruption rates, time pressure, and dynamic risk—conditions that amplify the relevance of attention management, cognitive workload, and adaptive coordination.

The purpose of this evaluation was not to estimate short-term incident reduction or direct behavioral change, which is difficult to attribute to brief messaging in complex sociotechnical systems. Instead, we focused on acceptance and behavioral intent as necessary preconditions for deployment: if messages are not accepted as reasonable and usable, they cannot plausibly support reliable behavior.

Participants first completed the profiling questionnaire (Section 3) to estimate personality tendencies and engagement. Based on the allocated PTS scope, each participant was presented with seven safety comments. After reading the comments, participants rated them anonymously on three 5-point Likert items: (i) perceived agreement/fit ("I can agree with the content"), (ii) practical usefulness ("It is helpful for medical safety activities"), and (iii) behavioral intent ("I intend to keep this in mind during my work").

Results and Practical Implications

Overall, participants evaluated the presented comments positively. As shown in Figure 5, the proportion of positive responses (Agree/Strongly agree) was 80% for agreement, 72% for usefulness in patient safety activities, and 74% for intention to apply the content at work.

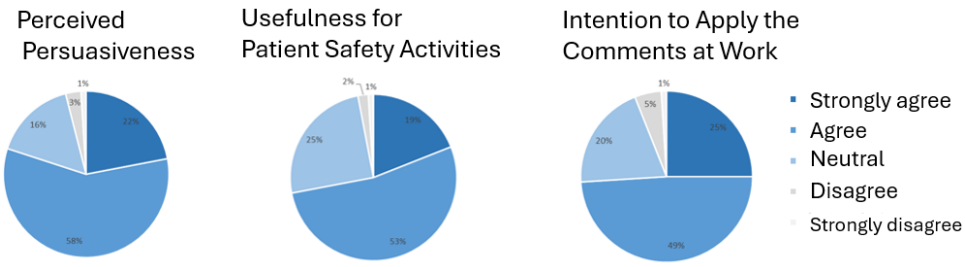

Figure 5: Survey results from nurses.

These results suggest that PTS-delivered comments achieve a high level of psychological acceptability and are perceived as practically relevant. From a human factors perspective, the engagement-based depth control appears to be a plausible strategy for managing intervention burden—supporting reflection and proactive safety involvement without imposing excessive cognitive load or provoking reactance.

Limitations

This evaluation relied on self-reported perceptions and intentions rather than direct behavioral measures or safety outcomes. The sample was limited to nurses in a single clinical context, and generalizability to other high-risk domains requires further study. Finally, although presenting multiple comments supports stable evaluation, the optimal frequency, timing, and dosage for sustained deployment—and the interaction with workload and attentional demands—remain open questions.

CONCLUSION

This study reframed safety micro-interventions as designable and operable artifacts tailored to workers' characteristics and state, rather than uniform messages delivered indiscriminately. This reframing is motivated by the observation that safety in high-risk work increasingly depends on human reliability under variability: sustained attention, adaptive judgment, and coordinated action under constraints.

We proposed PersonaTrace Scope (PTS), which decomposes personalization into two independently controllable axes: personality-based framing (regulating tone and rhetorical stance to support acceptability and responsibility ownership) and engagement-based depth control (regulating intervention demand to manage cognitive load and reduce reactance). This bounded design space enables operationally feasible personalization across 175 predefined scopes.

We also addressed scalable quality governance. Because exhaustive manual review of 175 scoped comment patterns is impractical, we positioned an LLM strictly as an evaluator—a quality gate—and introduced the Behavioral Trigger Index (BTI) as a multidimensional metric to enforce minimum action-triggering quality. BTI decomposes conditions relevant to behavior initiation and supports consistent quality control prior to deployment.

A field evaluation in nursing indicated that PTS-based comments were broadly acceptable and were associated with perceived usefulness and behavioral intent. These findings provide initial evidence that the framework can meet necessary preconditions for workplace use.

Future work should examine longitudinal links between scoped messaging and observable safety behaviors or performance indicators, including attention-sensitive outcomes (e.g., adherence to critical checks under interruption) and collaboration behaviors (e.g., speaking up, cross-checking). Further, research should explore how individual-level outputs can be aggregated to support team- and organizational-level monitoring and decision-making while preserving accountability and avoiding over-reliance on automation. Overall, PTS offers a practical pathway toward personalized, quality-assured safety support that is compatible with the realities of high-risk sociotechnical work.

REFERENCES

Abdul, A. F., & Ong, T., 2024. Real-World Outcomes of a Digital Behavioral Coaching Intervention to Improve Employee Health Status: Retrospective Observational Study. JMIR mHealth and uHealth, vol. 12, p. e50356.

Alslaity, A., Chan, G., & Orji, R., 2023. A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. Frontiers in Artificial Intelligence, vol. 6.

Ariki, C., & Okada, Y., 2006. A Study on the Worker's Personality and Human Error: Based on the Big Five Theory of Personality Psychology. In Proceedings of the 39th Annual Conference of the Japan Society for Safety Engineering (in Japanese).

Curcuruto, M. et al., 2024. Improving Workplace Safety Through Mindful Organizing: Participative Safety Self-Efficacy as a Mediational Link Between Collective Mindfulness and Employees' Safety Citizenship. Journal of Risk Research, vol. 27(1), pp. 85–107.

Daddy, D., Karta, S., & Intan, D. P., 2024. An Analytical Examination of Andragogical Principles in the Implementation of Adaptive Technology for Adult Education. Educational Technology, pp. 95–105.

de Jong, B., Jansen in de Wal, J., Cornelissen, F., & Peetsma, T., 2023. Investigating Transfer Motivation Profiles, Their Antecedents and Transfer of Training. Education Sciences, vol. 13(12), p. 1232.

Gallup, 2024. Gallup. [Online]. Available: https://www.gallup.com/q12-employee-engagement-survey/ [Accessed: 20 Jan 2026].

Gomes, O. T., Girão, F. B., Silva, H. M., & Andrade, M. V., 2025. Competency-based education in intensive care multiprofessional training: a scoping review. Critical Care Science, p. e20250385.

Jina, S. et al., 2024. Toward Tailoring Just-in-Time Adaptive Intervention Systems for Workplace Stress Reduction: Exploratory Analysis of Intervention Implementation. JMIR Mental Health, vol. 11, p. e48974.

Karve, Z. et al., 2025. New Doc on the Block: Scoping Review of AI Systems Delivering Motivational Interviewing for Health Behavior Change. Journal of Medical Internet Research, vol. 27, p. e78417.

Langdon, J. K. et al., 2021. Feasibility and acceptability of a digital health intervention to promote engagement in and adherence to medication for opioid use disorder. Journal of Substance Abuse Treatment.

London, M., Sessa, V. I., & Shelley, A. L., 2023. Developing Self-Awareness: Learning Processes for Self- and Interpersonal Growth. Annual Review of Organizational Psychology and Organizational Behavior, vol. 10, pp. 623–650.

Mantena, S., Johnson, A., & Oppezzo, M., 2025. Fine-tuning LLMs in behavioral psychology for scalable health coaching. npj Cardiovascular Health, vol. 2.

Min-Jeong, Y. et al., 2023. A Just-In-Time Adaptive Intervention (JITAI) for smoking cessation: Feasibility and acceptability findings. Addictive Behaviors, vol. 136, p. 107467.

Paterson, C. et al., 2024. Barriers and facilitators to implementing workplace interventions to promote mental health: qualitative evidence synthesis. Systematic Reviews, vol. 13(1), p. 152.

Perry, L. et al., 2024. Personality Feedback With Tailored Self-Care Recommendations Improves Self-Efficacy for Cancer Management: A Randomized Controlled Trial. Psycho-Oncology, vol. 33(11), p. e70023.

Persson, D., Bardram, J., & Bækgaard, P., 2024. Perceptions and effectiveness of episodic future thinking as digital micro-interventions based on mobile health technology. DIGITAL HEALTH, vol. 10.

Schleider et al., 2022. A randomized trial of online single-session interventions for adolescent depression during COVID-19. Nature Human Behaviour, vol. 6, pp. 258–268.

Selina, M., & David, E., 2025. LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour. International Journal of Human-Computer Studies, vol. 200.

Silla, I., Gajudo, J.-A., & Gracia, F. J., 2025. Safety in high-reliability organizations: The role of upward voice, team learning, and safety climate. Journal of Safety Research, vol. 93, pp. 55–65.

Timothy, M. D., Philip, P. M., Nathan, P. P., & Amber, Y. N., 2024. Harnessing the power of employee voice for individual and organizational effectiveness. Business Horizons, vol. 67(3), pp. 283–298.

Watanabe, Y., Nakayama, M., & Takemura, K., 2025. AI feedback and workplace social support in enhancing occupational self-efficacy: a randomized controlled trial in Japan. Scientific Reports, vol. 15.

Wing, L. et al., 2024. Echo 'Our' Voice? The influences of team members on the voice behavior of focal employees. Journal of Business Research, vol. 183.