

Integrated Decentering: Validating an Automated Measure of Decentering Expression Across Seven Independent Samples

Tamas Madl

`tamas.madl@ofai.at`

Austrian Research Institute for Artificial Intelligence

Sara Lazar

Massachusetts General Hospital

Article

Keywords: Decentering, Cognitive defusion, Mindfulness, Construct validity, Natural language processing, Psychological measurement

Posted Date: August 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10410851/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Integrated Decentering: Validating an Automated Measure of Decentering Expression Across Seven Independent Samples

Tamas Madl^{*1} and Sara W. Lazar^{†2}

¹Austrian Research Institute for Artificial Intelligence (OFAI); University of Vienna

²Massachusetts General Hospital

August 10, 2026

Abstract

Objectives. Decentering—relating to one’s thoughts and emotions as passing mental events rather than being fused with them—is difficult to measure at scale. We test whether Integrated Decentering (ID), an automated measure of the decentering expressed in short reflective text, shows construct validity as one fixed scoring model.

Methods. A purpose-trained language-model ensemble scored texts from seven independent samples (self-report, community, essay, diary, and online-argument samples, and two contemplative-practice cohorts). One cohort was a post-hoc secondary analysis of a registered randomized trial; ID was not a prespecified trial outcome. Aggregate-level analyses tested convergent, criterion, discriminant, incremental, reliability, and practice-linked validity, controlling for text length.

Results. ID converged with established decentering scales (EQ-D $r = .36$, 95% CI [.20, .50], $n = 124$; MPoDT $r = .41$ [.09, .65], $n = 35$); no prompted general-purpose model, frontier included, reproduced the EQ-D convergence. ID predicted empathic concern ($r = .40$ [.36, .44]), wise-reasoning search-for-resolution ($\beta = .18$ [.05, .30]), and value-behavior coherence ($r = .33$ [.24, .41]; $\Delta R^2 = .027$ beyond the Big Five). Predicted nulls held (adjacent wise-reasoning subscales, rumination, negative affect), and length-matched impersonal argument showed no positive discrimination (AUC = .45 [.40, .49]). Aggregates were stable over 12 months ($r = .69$ [.52, .81]). ID distinguished practitioners from matched non-practitioner controls (meditation $g = 1.83$ [1.08, 2.89]; yoga $g = 1.38$ [0.68, 2.29]); in a post-hoc non-randomized exposure analysis, reported sitting practice tracked 24-month gains ($g = .80$ [0.13, 1.56]). ID overlapped with an expert developmental score, as theory predicts, yet predicted decentering-relevant self-report beyond it.

Conclusions. One fixed scoring model transferred unchanged across all seven samples: ID offers a scalable measure of decentering expression in reflective text for contemplative science and other fields concerned with metacognitive distance.

Trial registration. ClinicalTrials.gov NCT02628548; registered 11 December 2015.

Keywords · Decentering · Cognitive defusion · Mindfulness · Construct validity · Natural language processing · Psychological measurement

^{*}Corresponding author: tamas.madl@ofai.at. ORCID: 0000-0001-6120-7855.

[†]ORCID: 0000-0003-1126-8363.

1 Introduction

1.1 Decentering and the problem of measuring it

A long contemplative and clinical lineage holds that suffering often stems less from thoughts and feelings themselves than from how one relates to them. Decentering is the metacognitive shift from being fused with mental contents—treating thoughts, emotions, and self-concepts as the self—to holding them as transient, constructed mental events. Bernstein, Hadash, and Fresco [Bernstein et al., 2015] frame this fusion-to-defusion process as a unitary metacognitive construct that spans contemplative and clinical traditions: decentered awareness in mindfulness-based cognitive therapy, cognitive defusion in acceptance and commitment therapy, and the broader mindfulness literature. The same shift is central to Shapiro and colleagues’ “reperceiving” account of the mechanisms of mindfulness [Shapiro et al., 2006]—one of several labels (decentering, cognitive defusion, reperceiving, self-distancing) under which the contemplative and clinical literatures have converged on a common metacognitive shift.

Self-report operationalizations of decentering mediate the effects of mindfulness-based interventions on clinical outcomes [Gu et al., 2015, van der Velden et al., 2015], and the same capacity is implicated well beyond contemplative practice—in emotion regulation, empathy, wise reasoning about conflict, and the perspective-taking that supports prosocial action. Experimentally, a decentered stance toward imagined stressful events reduces their affective impact [Lebois et al., 2015]. Decentering is thus a candidate mechanism of unusually broad reach, which makes its measurement consequential.

Two measurement channels dominate. The first is self-report: trait scales such as the Experiences Questionnaire (EQ) [Fresco et al., 2007] and the Metacognitive Processes of Decentering–Trait scale (MPoDT) [Hanley et al., 2020] (and, for the closely related ACT construct of cognitive fusion, the Cognitive Fusion Questionnaire [Gillanders et al., 2014]) are well validated and clinically useful. But they inherit the familiar limits of introspective report: sensitivity to demand characteristics, dependence on the respondent’s metacognitive accuracy, and a state-versus-trait distinction that varies by instrument and is easily overlooked (the Toronto Mindfulness Scale [Lau et al., 2006] targets state, the EQ and MPoDT trait). More acutely in contemplative populations, self-report also presupposes the very capacity it assesses: rating one’s own decentering requires meta-awareness of one’s fusion. Practitioners with heightened meta-awareness may under-rate their decentering precisely because they see their fusion more clearly, while novices over-rate theirs—a reference-group bias documented in mindfulness self-report [Grossman, 2008, Van Dam et al., 2018] that language-based measurement does not share in the same form, though (as we note below) it carries biases of its own.

The second channel, specific to contemplative research, is practice exposure: lifetime or in-trial practice hours used as a proxy for how much decentering an intervention has instilled. Yet exposure indexes delivery, not internalization. Home-practice–benefit correlations in mindfulness trials average only $r \approx .26$ [Parsons et al., 2017], dose–response meta-regressions find no robust practice effect on distress outcomes [Strohmaier, 2020], and in long-term meditators lifetime practice accounts for a modest, non-linear share of wellbeing variance [Bowles et al., 2022]. At matched exposure, some practitioners show durable change and others do not. Both capture internalized decentering only partially: one asks people to report a capacity they may not observe accurately, the other counts hours rather than their effect.

1.2 Language as a complementary measurement channel

Reflective language offers a third channel. How a person construes a charged situation in their own words—fused with a thought, emotion, or identity, or holding it as partial, transient, and constructed—carries information about decentering that neither a Likert rating nor an hour count contains [Lanning et al., 2018]. Naturalistic text reduces dependence on the respondent’s metacognitive accuracy and is less transparent to demand than a questionnaire item. Until recently the obstacle was scale: extracting this signal required expert human coding, as in the interview-based Measure of Awareness and Coping in Autobiographical Memory [Teasdale et al., 2002]—slow, costly, and hard to standardize. Language models now make it feasible to score reflective constructs from short free-text passages automatically and reproducibly, opening this channel to large, multi-sample validation.

Automated text methods have so far indexed the neighbouring construct of psychological self-distancing through surface lexical markers such as first-person-pronoun and present-tense rates [Nook et al., 2017, Kross and Ayduk, 2017]; ID instead scores decentering directly with a construct-trained model rather than relying on such markers, and is not reducible to them: a validated LIWC-22 distancing index is near zero with the criteria ID predicts in the two large samples (WASSA, Grossmann; Supplementary Materials S15). Prior language-model work in psychology has largely fit a supervised model to a single labeled outcome in a single dataset; we instead validate one fixed, construct-targeted scorer as a measurement instrument: across samples, against discriminant boundaries, and incrementally over alternative accounts.

1.3 Integrated Decentering

We introduce Integrated Decentering (ID), a measure that scores the decentering expressed in short reflective English text on a continuous scale. ID is a measurement method for the existing decentering construct [Bernstein et al., 2015], not a proposal for a new construct: the qualifier “integrated,” explained below, describes the level at which the measure is validated, not a new theoretical entity.

To make the construct concrete, consider two first-person responses to the sentence stem “When I am criticized...” (public-domain WUSCT-tradition corpus [Lanning et al., 2018], quoted verbatim):

More decentered: “I may feel hurt initially, but later try to honestly judge the validity of the criticism (difficult! [sic]).”

More fused: “I often agree with what has been said about me and become sad (once in a while mad).”

Both are comparable in length and name a negative reaction to the same stem; what differs is the stance toward that reaction—whether it is held as a passing mental event and examined, or taken as simply true of the self. These illustrate the construct, not ID’s scores: ID is validated on per-person aggregates, not single completions.

ID is produced by a purpose-built model ensemble (Section 2); here we note only the properties that bear on interpretation. ID is not a general-purpose chatbot prompted to rate passages; it is a fixed, purpose-trained scorer, fine-tuned with a contrastive objective to judge which of two reflective texts expresses more decentering. ID scores *texts*, *not people*: a single passage’s score is an estimate of the decentering expressed in that passage, and the validated unit is the aggregate (per person or condition for sentence completions; per text or event for essays and diaries), not the isolated single completion.

We call the measure *Integrated* Decentering because it indexes decentering not as a single act but as it has become woven into a person’s habitual meaning-making—metacognitive distance integrated into how experience is spontaneously construed and narrated. The training signal was built from reflective sentence-completion material in the Loevinger/WUSCT tradition and its adult-developmental extensions [Lanning et al., 2018, Madl and Cook-Greuter, 2026]; that material is a source of reflective text and contrastive structure, not the definition of the target. Accordingly, we do not treat ID as an automated developmental-stage rating, and we validate it directly against decentering criteria, not the developmental scoring from which its training pairs were drawn.

ID measures decentering as *expressed in language*, not decentering as lived. A contemplatively mature person may describe experience in plain, unmarked speech; a psychologically literate writer may produce decentered phrasing without the corresponding shift in perspective. The anti-shortcut training, text-length controls, and the null result on length-matched impersonal argument (Section 4.5) mitigate but cannot eliminate this gap. ID is best understood as one imperfect behavioral trace—a third channel alongside, and convergent with, the first-person perspective that self-report uniquely provides—not a replacement for it, nor a measure of contemplative attainment.

1.4 Validation logic

A new measure earns its interpretation not from any single correlation but from a coherent convergent-and-discriminant pattern across methods [Cronbach and Meehl, 1955, Campbell and Fiske, 1959]. We therefore draw on multiple sources of validity evidence for ID across seven heterogeneous samples, each assigned a prediction the construct should generate (Figure 1):

- **Convergent.** ID should correlate with established self-report decentering scales (EQ-D, MPoDT).
- **Nomological / criterion.** ID should predict theoretically adjacent outcomes—empathic concern, wise reasoning about conflict, and coherence between stated values and reported behavior—while remaining null on adjacent constructs the theory does not implicate.
- **Discriminant / boundary.** ID should show no positive discrimination in an impersonal, argumentative register where the construct does not apply, once text length is matched.
- **Practice-linked.** ID should be elevated in contemplative practitioners relative to controls, more so for the practice that directly trains decentering, and should track meditation exposure over time (the exposure test is post hoc; Section 3.2).
- **Incremental.** ID should carry information beyond text length and verbal ability, beyond broad personality, and beyond a human expert’s developmental score for the same protocols. This is the closest developmental benchmark, since it taps a maturity-relevant axis shared with ID yet is scored independently.

No single sample resolves the construct claim; the samples’ principal weaknesses differ across the design; and no single confound we could test can jointly produce the pattern—the argument is cumulative. The claim is correspondingly falsifiable: convergence indistinguishable from a generic verbal or maturity factor, criterion associations lacking the predicted specificity, or above-chance discrimination in the length-matched argumentative register would each count against it. Each alternative account or near-neighbor benchmark is addressed by a different cell of the design (Table 3): text length and verbal ability, broad personality, generic positivity and self-reflection, surface linguistic distancing, a human expert’s developmental score, and zero-shot LLM scoring (prompting a general-purpose model, without task-specific training).

The evidence is tiered: the convergent, nomological, behavioral/incremental, and discriminant tests form the core of the validation; the practice-linked results test sensitivity to contemplative exposure; and the neuroimaging analysis is reported as an exploratory extension. That extension shows ID recovers, automatically from text, a developmentally gated pattern of posterior-cingulate

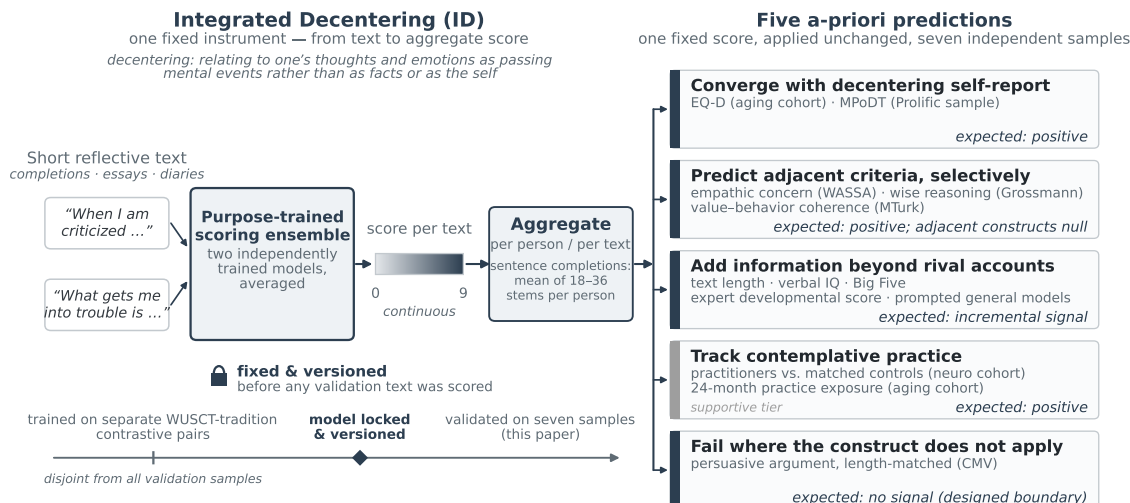


Figure 1: **Measurement pipeline and validation blueprint for Integrated Decentering.** One purpose-trained ensemble scores short reflective texts on a continuous 0–9 scale; validity is established at the aggregate level. The model was fixed before any validation text was scored and applied unchanged to all seven samples; the right panel shows the design’s five a-priori predictions, including predicted nulls and a predicted failure boundary (dark bars: core validation; grey: supportive tier). Outcomes are summarized in Table 2 and Figure 2.

connectivity previously reported for this cohort. Throughout, we delimit ID’s scope: a scalable research instrument for decentering expression, not a diagnostic, neural marker, or developmental-stage measure.

Practically, because ID scores text that studies already collect, any trial or cohort with open-ended reflections, practice diaries, or exit interviews can rescore that existing text for a decentering measure at no additional participant burden—an applied payoff we return to in the Discussion.

2 The Integrated Decentering scorer

2.1 What the scorer estimates

ID is a continuous scorer of the decentering expressed in short reflective English text. Given a passage, it returns a value on a 0–9 scale, where higher scores indicate stronger textual evidence of stepping back from one’s thoughts, affects, identities, or perspectives and holding them as transient, constructed events. The scorer estimates a property of the *text*, not a stable attribute of the writer; aggregates—per-person or per-condition means for sentence completions, and per-text or per-event scores for essays and diaries—are the unit at which validity is established and at which the analyses below are reported. Full architectural and training detail is given in Supplementary Materials S1; we summarize here only what is needed to interpret the results.

2.2 Architecture and training

Conceptually, the scoring model was trained much as one might train a human rater. It was shown many thousands of pairs of reflective texts—in each pair, one expressing more decentering than the other—and learned to judge which of the two expresses more. Many of these pairs were

matched on developmental stage and on text length, so the model could not tell the two apart by overall maturity or by verbosity and had to rely on the decentering contrast itself (we call these “anti-shortcut” pairs). ID still shares a maturity-relevant axis with expert developmental scoring but does not reduce to it, as we quantify later. The final measure averages the judgements of two independently trained models, analogous to averaging two trained raters. Its parameters were fixed and versioned before any validation text was scored; no dataset-specific tuning, prompting, or calibration occurred at any point.

The technical implementation—an equal-weight ensemble of two independently trained low-rank (LoRA) adapters over open-weight Qwen base models—is detailed in Supplementary Materials S1.

Rather than regressing text onto a developmental-stage label, each member was trained with a **contrastive Bradley–Terry objective** [Bradley and Terry, 1952] on pairs of passages, predicting which of two texts expresses more decentering. The training signal was derived from reflective sentence-completion material in the Loevinger/WUSCT tradition and its adult-developmental extensions [Lanning et al., 2018, Madl and Cook-Greuter, 2026], using a decentering-aligned indicator from that upstream work as a deliberately imperfect, theory-motivated proxy. The length-matched pairs were of three kinds; the up-weighted **within-stage pairs** (same developmental stage, different decentering proxy) suppress “stage shortcut” learning, so the scorer tracks reflective distance rather than a coarse developmental bucket (Supplementary Materials S1). ID and expert developmental scoring nonetheless share a maturity-relevant axis—the subject–object differentiation that decentering and ego development are theorized to hold in common—so the two agree substantially; this expected overlap reflects that shared axis and same-protocol scoring rather than stage recovery, and is not by itself evidence of validity (Section 3.2). The construct claim therefore rests not on the size of that agreement but on differential prediction—the non-developmental criteria below (empathy, wise reasoning, value–behavior coherence), and the fact that ID still predicts decentering-relevant self-report after the expert score is accounted for (Section 4.5).

2.3 Optional eligibility filtering

Decentering is meaningfully scoreable only when a passage contains first-person or self-relevant material the writer could identify with—a belief, value, feeling, identity, or perspective. Purely factual text carries no such signal; any score on it is dominated by length. For heterogeneous inputs we therefore use two pre-specifiable eligibility filters that operationalize this precondition: a deterministic *mechanical* filter (first-person, belief, or affect markers) and a more permissive *LLM* filter (a separate model used only to judge eligibility, not the ID scorer; Supplementary Materials S2). Filtering also removes unscoreable, length-only completions from short sentence-completion data; this matters most in small samples. We report the unfiltered view alongside as a sensitivity check and treat filtered estimates as primary only for the mixed-register essay corpus. Filter views are subset operations on the same per-text scores (Supplementary Materials S2). Because the sentence-completion stems are shared across participants, we also verified that scoring the completion alone rather than the full stem-plus-completion leaves convergent estimates materially unchanged, ruling out a shared-stem artifact (Supplementary Materials S4).

2.4 Scope and boundary conditions

ID is validated for written, reflective English text of roughly sentence-to-paragraph length; it is strongest for self-relevant reflective material and most defensible when aggregated. It has not been validated for other languages or non-reflective registers, and should not be used for decisions about identifiable individuals or to infer a person’s mindfulness, wisdom, mental health, or developmental

stage. One boundary is itself a validation target: on impersonal text that is neither self-relevant nor emotionally engaged (e.g. persuasive argument about an external topic) the scorer can behave as a length proxy (Section 4.5), so it should not be applied to such content as if it measured decentering.

The training material is disjoint from every validation sample at the participant and response level: none of the validation data evaluated here entered training (the sentence-completion prompts are the standard shared format, but no training individual or response appears in the validation samples). A distinct concern is *pretraining* exposure: some validation samples (WASSA, CMV, MTurk) are public and could occur in the base models’ pretraining data. Three observations bound this concern: the task is judging a passage, not recalling it, and no criterion labels are paired with text in any public release in scoreable form; the untuned base ensemble—with identical pretraining—carries no convergent signal (Supplementary Materials S15); and the convergent, reliability, and practice-linked evidence comes from private clinical cohorts that are not part of any public corpus.

3 Validation datasets and analytic strategy

3.1 Participants and measures, organized by validation role

We validate ID across seven independent samples (Table 1), each carrying one or more validation roles. The samples are deliberately heterogeneous in population, text genre, and criterion type, and were not collected for ID; several predate the scorer entirely. Brief descriptions follow; full per-sample procedures, measures, and sample characteristics are in the Supplementary Materials.

Convergent. The **MGH aging** cohort (below) provides the larger, better-powered convergent anchor: the 11-item Experiences Questionnaire decentering subscale (EQ-D) [Fresco et al., 2007], with $n = 124$ holding both ID and EQ-D. The **MPoDT** sample, collected by the authors via Prolific, comprises 35 adults who completed sentence-completion stems and the Metacognitive Processes of Decentering–Trait scale (MPoDT) [Hanley et al., 2020], an established text-relevant decentering self-report; it serves as a smaller corroborating anchor whose subscale structure additionally probes construct coverage (Section 4.2).

Nomological / criterion. **WASSA** [Buechel et al., 2018] is a corpus of 1,860 first-person essays written in response to news articles, each rated for empathic concern and personal distress; decentering should track the compassionate (non-fused) component of empathy. The **Grossmann** conflict-diary sample [Grossmann et al., 2021] contains 424 first-person accounts of interpersonal conflict (298 participants), each coded by trained raters on five wise-reasoning subscales [Brienza et al., 2018]; decentering should predict search-for-resolution and perspective-taking but not the three adjacent subscales (compromise, recognition of limits, expectation of change). The **MTurk** community sample [Fournier et al., 2022] comprises 446 middle-aged adults with sentence completions, the Big Five Inventory–2 (BFI-2), and a value–behavior coherence index computed across three close relationships.

Discriminant / boundary. The **CMV** corpus [Tan et al., 2016] pairs, for each Reddit r/ChangeMyView post, one argument that changed the poster’s view with one that did not (4,311 pairs). Persuasive success in this impersonal argumentative register should not reflect decentering once text length is matched.

Table 1: Validation samples organized by the facet of construct validity they address. ID-relevant text genre and the criterion against which ID is tested are shown for each.

Validity role	Sample (N)	Text genre	Criterion
Convergent	MGH aging (124)	Sentence completion	EQ-D decentering [Fresco et al., 2007]
	MPoDT (35)	Sentence completion	MPoDT trait decentering [Hanley et al., 2020]
Nomological / criterion	WASSA (1,860)	Empathy essays	Empathic concern [Buechel et al., 2018]
	Grossmann (424)	Conflict diaries	Wise-reasoning subscales [Grossmann et al., 2021, Brienza et al., 2018]
	MTurk (446)	Sentence completion	Value-behavior coherence [Fournier et al., 2022]
Discriminant / boundary	CMV (4,311 pairs)	Online argument	Persuasion, length-matched [Tan et al., 2016]
Practice-linked	MGH neuro (45)	Sentence completion	Practitioner vs. control [Gard et al., 2014, Singleton et al., 2021]
	MGH aging (143)	Sentence completion	Meditation dose; test-retest
Incremental	MTurk (446)	Sentence completion	Beyond the Big Five (BFI-2)
	MGH neuro (45)	Sentence completion	Beyond verbal IQ (AMNART); expert TWS benchmark

Notes: TWS = a human expert’s Total Weighted Score, a cumulative expert score of ego development for the same protocols, scored independently of the language model (see Section 3.2).

Practice-linked and incremental. The **MGH neuroimaging** cohort [Gard et al., 2014, Gard et al., 2015, Singleton et al., 2021] is a secondary analysis of 45 adults—meditation practitioners ($n = 14$), yoga practitioners ($n = 16$), and matched controls ($n = 15$)—with sentence completions, a behavioral battery (incl. Perceived Stress Scale (PSS), Quality of Life Inventory (QOLI), Self-Compassion Scale (SCS), American National Adult Reading Test (AMNART) verbal IQ), resting-state fMRI (posterior cingulate cortex [PCC] seed connectivity), and an expert human developmental rating of the same protocols. The **MGH aging** cohort is a post-hoc secondary analysis of a randomized trial of meditation-naïve older adults [Sevinc et al., 2021], prospectively registered on 11 December 2015 (ClinicalTrials.gov NCT02628548). The registry records 175 enrolled in the overall study, the primary report records 146 randomized (72 mindfulness, 74 cognitive-fitness control), and the present secondary dataset contains sentence-completion text for 143 participants at one or more analyzed waves; these are different recruitment, randomization, and data-availability denominators (full accounting in Supplementary Materials S10). Sentence completions were available at baseline, 12, and 24 months, with practice follow-up available for a subset. This single cohort provides three of the results reported below—the EQ-D convergent correlation, the 12-month retest stability, and the 24-month dose-response—which are therefore not independent: they are separate analyses (different waves and subsamples) of one study. Neither dataset was collected with ID in mind, and neither set of ID analyses was pre-registered; ID was not a prespecified trial outcome.

3.2 Analytic strategy

All analyses use per-person (or, for the essay/argument corpora, per-text) ID aggregates. Because ID is text-derived, every model controls for verbosity (log word count) and, where relevant, language covariates; convergent and criterion associations are Pearson or partial correlations net of these controls. Group differences use Welch’s t with Hedges’ g (a small-sample-corrected analogue of Cohen’s d), reported with exact noncentral- t 95% confidence intervals. Criterion and connectivity models use ordinary least squares with HC3 heteroskedasticity-robust standard errors; the Grossmann diary models, which contain multiple events per participant, use subject-clustered standard errors. Incremental validity is assessed by hierarchical ΔR^2 (e.g., ID added to a model already containing the Big Five). The discriminant boundary uses paired and length-matched area-under-the-curve (AUC). Headline effects are reported with 95% confidence intervals; secondary, filtered, and covariate-adjusted estimates are reported with exact p -values (reproducible script `25_headline_cis.py`).

ID is also compared against a human expert’s Total Weighted Score (TWS)—a cumulative expert score of ego development—for the same sentence-completion protocols, scored independently of the language model and previously used on the neuroimaging cohort by Singleton et al. [Singleton et al., 2021]. A substantial ID–TWS correlation is expected for two compounding reasons: the two tap a shared, maturity-relevant axis (the subject–object differentiation at the core of ego development [Kegan, 1982, Cook-Greuter, 1999]), and both are scored from the same protocols, so shared text properties (e.g., articulateness) inflate their agreement; the scorer’s contrastive training nonetheless discourages stage-bucket shortcuts (Section 2). TWS is a *cumulative* score, not a developmental stage: stage is assigned to a protocol by threshold (cut-off) rules. Because the same total can arise from different response profiles, the score-to-stage map is many-to-one, so agreement with TWS is uninformative about whether ID recovers discrete stage. Consistent with this, ID correlates with discrete stage well below TWS ($r = .69$ versus $r = .95$); squaring these, stage accounts for about 48% of ID’s variance against roughly 90% of TWS’s, so the within-stage training attenuated, rather than eliminated, the stage association (Supplementary Materials S9). We test whether ID carries information beyond TWS using Williams’ t for the difference between two dependent correlations sharing a variable ($df = n - 3$; eq. 7 in [Steiger, 1980]) and joint models partitioning their unique contributions (ΔR^2); we report the human-coded TWS throughout the main text and a model-derived maturity score only as a robustness check (Supplementary Materials S14).

For theory-specified tests with a priori directional hypotheses (convergent correlations, known-groups separation, the criterion associations, the discriminant boundary), we use $\alpha = .05$ uncorrected; for exploratory sets of correlations within a sample we apply Benjamini–Hochberg false-discovery-rate (FDR) control. For completeness, the primary directional family (convergent, criterion, and known-groups tests, $m = 7$) also survives Benjamini–Hochberg FDR control at $q = .05$, with a largest surviving p of .007 (Grossmann Search-for-Resolution; script `28_robustness_followups.py`). The exploratory neural moderation is the one exception: its regions and developmental split are hypothesis-constrained from prior work on this cohort, and it is reported uncorrected and explicitly flagged as not surviving correction. Every regression effect was checked for robustness with leave-one-out and Cook’s-distance diagnostics, and every filtered analysis against the unfiltered view; both are reported in the supplement. Effect sizes are interpreted against individual-differences norms in which $r \approx .10$, $.20$, and $.30$ are small, typical, and relatively large, with $r \geq .50$ rare [Gignac and Szodorai, 2016]. Verified, reproducible code for every number in this paper is provided (see Data and Code Availability).

Analysis provenance. Because neither contemplative cohort was collected for ID and none of these analyses was preregistered, we distinguish analyses by evidential status. *Theory-guided primary tests*—the convergent correlations, the criterion associations (empathy, wise reasoning, value–behavior coherence), the known-groups separation, and the length-matched discriminant boundary—each test an a priori directional hypothesis. *Pre-specifiable filter logic*: the two eligibility filters operationalize a stated scoreability precondition and are fixed by register rather than tuned to outcomes (Section 2), with the unfiltered view always reported alongside. *Robustness and sensitivity checks*—filter-view comparisons, verbosity and demographic adjustments, cross-validated increments, and leave-one-out diagnostics—probe whether a primary result depends on an analytic choice. *Post-hoc exposure analyses*: the 24-month dose–response is non-randomized and conditioned on who was reached and chose to practice, and is labelled as such throughout. Within it we treat the categorical any-sitting-practice contrast ($g = 0.80$) as the primary exposure estimate, with the continuous frequency association ($\rho = .36$) and the joint practice-vs-arm model as robustness checks on the same exposure rather than independent tests. *Exploratory neural analyses*: the connectivity moderation is hypothesis-constrained from prior work on this cohort but reported uncorrected and flagged as not surviving correction. We report all of these, not only the confirmatory subset.

4 Results

Results are organized by validity type rather than by sample, each subsection assembling evidence across whatever samples bear on it. Table 2 and Figure 2 summarize every result on one effect-size axis. The argument is cumulative; no single effect carries the construct claim.

4.1 Measurement behavior and reliability

ID scores were approximately normally distributed in every sample and correlated modestly with text length (Spearman $\rho \approx .2$ – $.3$ across samples). Part of this is expected on substantive, not only psychometric, grounds: holding a thought or feeling as a transient, examinable mental event tends to take more words than asserting it as simple fact, so decentering plausibly co-occurs with somewhat fuller elaboration of inner experience rather than length being purely a nuisance. We nonetheless control for verbosity (log word count) throughout; this is conservative—it removes construct-relevant along with nuisance variance—so the criterion associations that survive it are, if anything, understated.

A single completion is a weak index of the person (single-stem intraclass correlation $ICC(1, 1) \approx .05$ – $.09$), as expected for one short stem, but averaging over a participant’s stems (mean 32, 18, and 36 per person) yielded reliable aggregate scores ($ICC(1, k) = .75$ in the neuroimaging cohort, $.65$ in the MPoDT sample, and $.66$ per person-wave in the aging cohort; split-half Spearman–Brown $r = .73$; single-stem reliabilities in Supplementary Materials S1). ID is therefore analyzed and reported at the per-person or per-condition level in these sentence-completion cohorts. The two ensemble members agreed on training-format text ($r \approx .93$) but diverged enough across the validation battery to justify averaging (Supplementary Materials S1).

In the randomized aging cohort, the active-control arm (assigned games and puzzles, not meditation) gives a 12-month stability estimate: Pearson $r = .69$ (95% CI $[.52, .81]$, $n = 56$), comparable at 24 months and in the range of established trait measures. Because some control participants practiced on their own (Section 4.6), this reflects the stability of a partly practice-exposed sample rather than a pure intervention-free estimate. This stability is the first evidence that ID indexes a *consolidated* disposition rather than a momentary state: the same individuals received similar scores

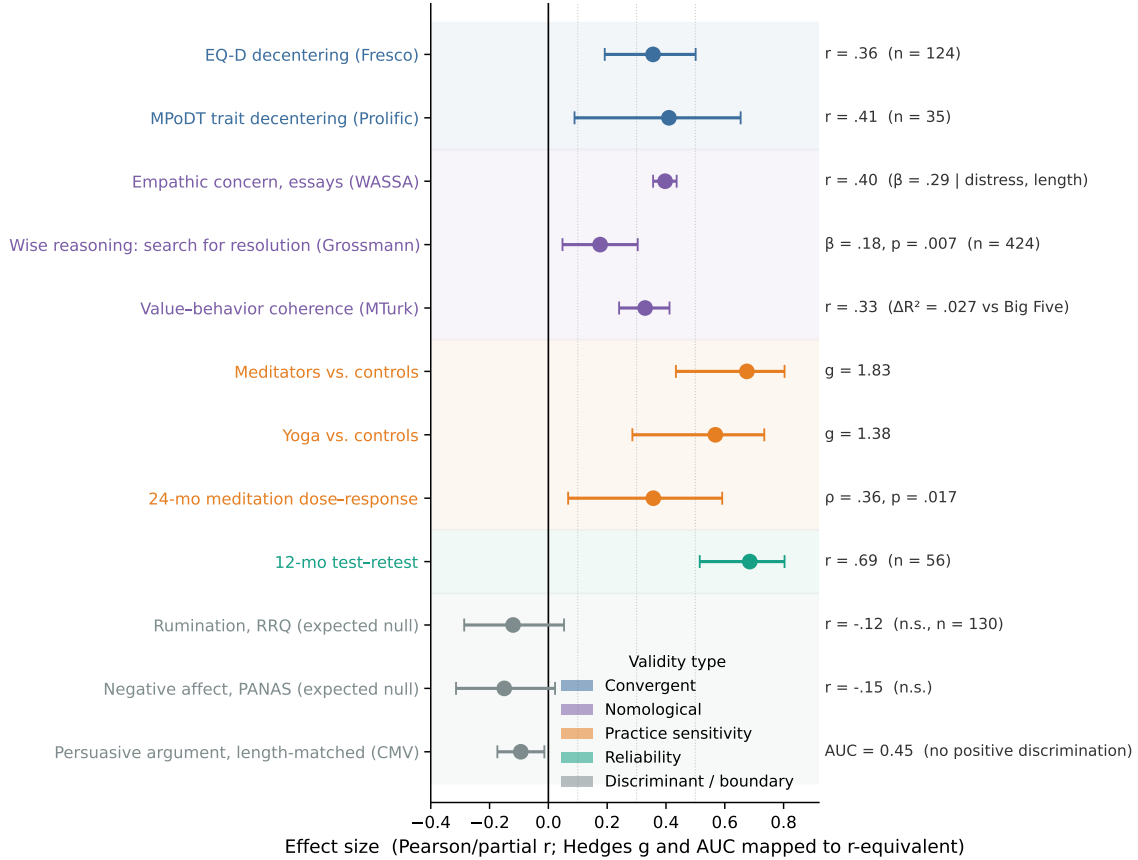


Figure 2: **Validity evidence across seven samples for Integrated Decentering.** Rows are placed on an approximate common effect-size axis for visual comparison only: Pearson/partial r is shown directly, while Hedges' g and AUC are converted to approximate r -equivalents through Cohen's d (AUC via the equal-variance normal approximation). Native statistics are printed at right and should be used for interpretation. Bars are 95% CIs computed from the reported n . The construct is supported where predicted (convergent, nomological, practice-linked), is null where it should be (rumination, negative affect), and shows no positive discrimination in the off-construct, impersonal argumentative register (CMV).

Table 2: Sources of validity evidence for Integrated Decentering. Headline effects by validity type; full statistics, covariates, and robustness checks are in the cited Results subsections and Supplementary Materials.

Validity type	Test	Effect	§
<i>Core validation</i>			
Convergent	EQ-D decentering (aging, $n=124$)	$r = .36$	4.2
	MPoDT trait decentering (Prolific, $n=35$)	$r = .41$	4.2
Nomological	Empathic concern (WASSA, $n=1,705$)	$r = .40$; $\beta = .29$	4.3
	Wise reasoning: search-for-resolution (Grossmann)	$\beta = .18$, $p=.007$	4.3
	adjacent wise-reasoning subscales	null (3/3)	4.3
Behavioral / incremental	Value-behavior coherence (MTurk)	$r = .33$	4.4
	beyond the Big Five	$\Delta R^2 = .027$, $p=.0001$	4.4
Discriminant / boundary	Persuasion, length-matched (CMV)	AUC = .45 (no positive discrimination)	4.5
	Rumination, negative affect (divergent)	null ($r = -.12$, $-.15$)	4.3
	Beyond an expert TWS score: tracks	Shared axis + same protocol; not reducible	4.5
	SCS more closely, QOLI sign divergence, predicts decentering-relevant self-report net of maturity		
Reliability	12-mo test-retest	$r = .69$	4.1
<i>Supportive: practice-linked</i>			
Practice sensitivity	Meditators / yoga vs. controls	$g = 1.83 / 1.38$	4.6
	24-mo meditation dose-response (post-hoc)	$g = .80$; $\rho = .36$	4.6
<i>Exploratory extension</i>			
Neural	PCC-IFC (whole sample); PCC-TPJ (gated)	$r = .54$; $\times$ stage $p=.046$	S13

Evidence tiers: *core* = the convergent, nomological, behavioral/incremental, discriminant, and reliability tests on which the construct claim rests; *supportive* = practice-linked evidence (cross-sectional known-groups and post-hoc exposure); *exploratory* = the non-independent neural reanalysis. Raw headline effects are shown for readability and are reported with 95% confidence intervals; secondary and covariate-adjusted estimates are reported with exact p -values in the cited subsections, and conclusions are unchanged under verbosity controls.

a year apart, even as the measure remained responsive to sustained practice (Section 4.6). Aggregate ID is also robust to the choice of eligibility filter; filter views change convergent correlations by $\leq .06$ and never reverse a conclusion (Supplementary Materials S2, S14).

4.2 Convergent validity

A valid decentering measure should converge with established self-report decentering scales; ID converged with both. The larger, better-powered anchor is the EQ-D: in the aging cohort, ID correlated with the EQ-D subscale at $r = .36$ (95% CI [.20, .50], $n = 124$, $p < .001$), holding beyond verbosity on the scoreable view (partial $r = .33$) and stable across filter views ($r = .30$ –.33). The smaller MPoDT sample corroborates: per-person ID correlated with the MPoDT trait-decentering composite at $r = .41$ (95% CI [.09, .65]) across all stems, rising to $r = .47$ with the mechanical eligibility filter, with the strongest associations on the subscales most squarely about decentering—Meta-awareness ($r = .49$) and Disidentification ($r = .43$)—and a weaker association with Reduced Reactivity ($r = .17$); its length-controlled partial ($r = .32$, $n = 35$, $p = .06$) is filter-dependent,

attenuating when unscorable, length-only completions are retained (Supplementary Materials S4). Pooled across the two anchors (inverse-variance Fisher- z), the convergent correlation is $r = .37$ (95% CI [.22, .50], combined $n = 159$; script `27_convergent_pooled.py`).

These magnitudes—moderate-to-large by individual-differences norms [Gignac and Szodorai, 2016]—indicate reliable but non-redundant convergence: ID taps the same construct while leaving most of its variance unshared, related but distinct rather than a substitute (Supplementary Materials S4). The observed values also understate the underlying convergence: correcting for measurement unreliability in both instruments (aggregate ID reliability .66–.75, Section 4.1; EQ-D internal consistency .83–.90 [Fresco et al., 2007]) places the disattenuated EQ-D correlation at roughly .44–.49 (Supplementary Materials S4); we report observed correlations throughout. Cross-method convergence of this kind is bounded by shared method variance and reliability. Correlations of this magnitude are typical when one construct is measured across different expression modalities (self-report versus text), much as self-informant agreement on personality has a ceiling well below unity [Connelly and Ones, 2010], and the disattenuated estimate approaches typical between-questionnaire convergence.

4.3 Nomological net: empathy and wise reasoning

Empathic concern (WASSA). Decentering should support the compassionate, non-fused component of empathy. In 1,860 first-person essays, ID correlated with rated empathic concern at $r = .35$ (raw text); under the pre-specified eligibility filtering for this mixed-register corpus, $r = .40$ (95% CI [.36, .44], $n = 1,705$, 92% retained). The association survived adjustment for personal distress ($r = .30$) and a regression controlling distress and log length ($\beta = .29$, $z = 13.0$). ID thus predicts empathic concern, the compassionate (non-fused) component, over and above personal distress, as theory predicts (Supplementary Materials S5).

Wise reasoning about conflict (Grossmann). If ID captures decentering rather than generic wisdom, it should predict the resolution-seeking facet of wise reasoning while leaving the adjacent subscales the theory does not implicate untouched. In 424 conflict-diary events (298 participants, subject-clustered SEs, controlling event length and survey wave), ID predicted the primary wise-reasoning subscale, Search-for-Resolution, at $\beta = .18$ (95% CI [.05, .30], $p = .007$), with Perspective-Taking marginal and secondary ($\beta = .10$, $p = .07$). The three adjacent subscales—Compromise, Recognition of Limits, and Expectation of Change—were all null ($\beta = .05, .02, -.01$; all $p > .4$), with 95% confidence intervals excluding standardized effects beyond $|\beta| = .16$: any undetected association there is at most small, consistent with discriminant specificity rather than proof of exactly zero effect (Supplementary Materials S6). We tested these coefficient differences directly with a subject-clustered bootstrap. ID’s Search-for-Resolution coefficient exceeded the mean of the three adjacent subscales by $\Delta\beta = .16$ (95% CI [.04, .27]) and exceeded each individually (one-sided $p \leq .04$), while remaining statistically indistinguishable from Perspective-Taking ($\Delta\beta = .07$, n.s.), the other theory-implicated subscale. The pattern supports a decentering-specific reading rather than a generic reflectiveness factor (script `28_robustness_followups.py`).

Adaptive reflection, not rumination. A decentering measure should track adaptive self-examination but not repetitive negative thought, its closest discriminant neighbor. At baseline in the aging cohort, ID correlated with Rumination–Reflection Questionnaire (RRQ) self-reflection ($r = .42$), San Diego Wisdom Scale (SD-WISE) insight ($r = .47$), positive affect ($r = .39$), and overall mindfulness (Five Facet Mindfulness Questionnaire, FFMQ; $r = .31$; all $p < .001$, $n \approx 130$), but was null with RRQ rumination ($r = -.12$, 95% CI [–.29, .05]) and negative affect ($r = -.15$,

$[-.31, .03]$), the intervals bounding both associations below $|r| = .31$. The dissociation (reflection positive, rumination null) is what decentering predicts and not what a relabeling of verbosity or distress would produce (Supplementary Materials S11); a separate role-play corpus shows the same reduced-reactivity pattern (ECSC/COPING; Supplementary Materials S12).

4.4 Behavioral and incremental validity

Beyond self-report and rated text, ID was related to a behaviorally anchored outcome. In the MTurk community sample ($n = 446$), ID predicted value-behavior coherence (VBC)—the within-person alignment between stated interpersonal values and reported behavior with three close others, from separate instruments—at partial $r = .33$ (95% CI $[.24, .41]$; raw $r = .37$; controlling word count and English fluency; robust across eligibility-filter views, $.32-.34$), decomposing into normative ($r = .27$) and distinctive ($r = .20$) components. ID correlated with Agreeableness ($r = .41$) and Openness ($r = .29$), as predicted; the stricter test is whether ID predicts VBC *beyond* broad personality. Adding ID to a model with all five Big Five domains increased explained VBC variance by $\Delta R^2 = .027$ ($F(1, 408) = 15.5$, $p = .0001$; $\beta_{ID} = .18$), and the increment held out-of-sample (10-fold cross-validated $\Delta R^2 = .022$). ID captures value-behavior alignment that broad personality dimensions do not.

As a corroborative test, ID correlated with coherence in autobiographical narratives at raw $r = .33$, but this attenuated to partial $r = .12$ ($p = .011$) once narrative word count was controlled, indicating that shared verbal productivity carries most of the cross-task signal; we therefore treat narrative coherence as directionally consistent rather than as a primary effect (Supplementary Materials S7).

4.5 Discriminant validity and construct boundaries

A construct measure must also fail where the construct does not apply, and must add information beyond cheaper or more general explanations.

An off-construct register (CMV). A demanding test of a measure is where it should fail: a length or articulateness proxy would win an impersonal persuasion task, but a decentering measure should not. In the ChangeMyView corpus, persuasive success is exactly such an off-construct property. ID’s paired AUC was a modest $.55$ (chance = $.50$), but this was confounded with text length (length alone gave AUC = $.62$). After matching arguments to within 10% on length, ID showed no positive discrimination (AUC = $.45$, 95% CI $[.40, .49]$): the apparent signal was length—the boundary the construct predicts (Supplementary Materials S8).

Beyond text length and verbal ability. ID correlates modestly with verbal sophistication, as any language measure must, but the criterion associations above all survive verbosity control, and the length-matched CMV boundary and the incremental-beyond-Big-Five result are both hard to reconcile with a pure verbal-fluency account: a length/articulateness proxy would have won the (length-driven) CMV task and would not add over personality. Directly, in the neuroimaging cohort, premorbid verbal IQ (AMNART-EVIQ) correlated with ID only modestly ($r = .33$, $n = 33$), and adding it to that cohort’s criterion models shifted the ID coefficient by $\leq .02$ on every outcome while AMNART itself predicted none of them (all $p > .29$; Supplementary Materials S9).

Nor is ID a demographic proxy: it related only modestly to age and sex ($r \approx .13$) and negligibly to education ($r = .01$), and the criterion effects survived adjustment for age, sex, and education

(MTurk value-behavior coherence $.33 \rightarrow .32$; the neuroimaging stress and self-compassion effects unchanged; Supplementary Materials S3).

Within a single sample, a supervised lexical model fit on the target can match any fixed instrument—that is what supervised fitting does. The distinctive property of a measurement instrument is transfer. The fixed ID score—never trained on these labels—matched a cross-validated TF-IDF+Ridge model (a standard text classifier built from word-frequency features) fit directly on WASSA empathy ($r = .35$ vs. $.36$) and exceeded it on the Grossmann criterion ($.20$ vs. $.01$). Each lexical model was sample- and target-specific, and on Grossmann fell below even the length baseline when required to generalize across writers; the same single ID model remained above the length baseline on both criteria and transferred across all samples with its discriminant boundaries intact.

On the better-powered convergent test, prompting a general-purpose model did not substitute for the trained scorer. Six models prompted with the decentering definition alone (zero-shot)—ID’s own untuned base models, two mid-tier models, and the two most capable general-purpose systems available at the time of analysis (model identifiers and query dates in Supplementary Materials S15)—gave $r \leq .14$ on EQ-D (the most capable $\leq .08$), versus ID’s $r = .31$ on the matched wave ($r = .36$ grand-average). Neither enabling reasoning nor few-shot prompting with anchored exemplars (adding scored examples) moved the most capable systems off this floor (few-shot $r = .06$ and $.08$). The untuned base ensemble (ID’s identical architecture before tuning) scored only $.07$, so the convergent signal came from the contrastive fine-tuning rather than the base model or its size. The small MPoDT sample ($n = 35$) is less able to separate models, so the better-powered EQ-D is the reliable comparison (Supplementary Materials S15).

Beyond a human expert’s developmental score. A close developmental benchmark for ID is a human expert’s Total Weighted Score (TWS) for the same sentence-completion protocols. As set out in Methods (Section 3.2), a substantial ID–TWS correlation is expected (from a shared, maturity-relevant axis and same-protocol scoring), so the relevant question is not its size but whether ID carries decentering-relevant information beyond it. In the neuroimaging cohort, ID correlated with TWS at $r = .79$ (95% CI $[.65, .88]$), sharing about 62% of its variance ($r^2 \approx .62$) and leaving roughly 38% unshared. We use TWS not as independent evidence for ID but as a benchmark for differential prediction. Indeed, ID predicted three decentering-relevant self-reports (the FFMQ meta-awareness and reduced-reactivity facets, and self-compassion) with little attenuation after partialling out TWS, whereas TWS tracked them only weakly (facet-level; see Supplementary Materials S9).

ID tracked self-compassion more closely than TWS ($r_{ID} = +.49$ vs. $r_{TWS} = +.27$; Williams’ $t(39) = 2.46$, $p = .019$). For quality of life the two scores diverged in sign ($r_{ID} = +.12$ vs. $r_{TWS} = -.12$; $t(40) = 2.59$, $p = .013$); since neither correlation was individually reliable, this is best read as a difference in pattern rather than a magnitude effect. Both differences held under Benjamini–Hochberg correction across the three a-priori psychological outcomes ($q = .028$). For perceived stress the two measures were directionally aligned ($r_{ID} = -.44$, $r_{TWS} = -.33$), and in joint regressions adjusting for practice group ID added more unique variance beyond TWS than the reverse for both stress ($\Delta R^2 = .069$ vs. $.013$) and self-compassion ($.112$ vs. $.072$); the dependent-correlation difference for stress did not reach significance in this small cohort. The ID associations also persisted controlling practice group, age, and verbosity (full test table, including the exploratory neural outcomes, in Supplementary Materials S9; a model-derived maturity score gives the same pattern, S14).

These comparisons rest on a single small cohort ($n = 42$ – 45 per outcome), so they are suggestive

rather than decisive. Together they support the conclusion that ID shares a maturity-relevant axis with a broad expert developmental score but indexes a specific, decentering-relevant capacity not reducible to it.

4.6 Sensitivity to contemplative practice

Known-groups specificity (neuroimaging cohort). ID distinguished practitioners from matched non-practitioner controls in the theory-predicted pattern: meditation practitioners exceeded controls at Hedges’ $g = 1.83$ (95% CI [1.08, 2.89]) and yoga practitioners at $g = 1.38$ ([0.68, 2.29]) (Figure 3B); the meditation effect was numerically larger, consistent with seated meditation more directly training decentering, though the two practitioner groups are small ($n = 14$ and 16) and the difference is not statistically distinguishable. These magnitudes reflect a cross-sectional extreme-groups comparison—long-term practitioners versus non-practitioner controls—and should not be read as expected intervention effects, which are far smaller.

Among meditators, ID exceeded *every* FFMQ facet ($g = 1.83$ vs. 1.56 next-best); for yoga practitioners ID ($g = 1.38$) was comparable to the strongest facet (Observe, $g = 1.51$) and exceeded the rest (Figure 3A). Only the two decentering-relevant facets (Observe, Non-reactivity) discriminated strongly; the three distal facets were near zero in meditators and modest in yoga ($g \leq .55$). Within meditators (standardized to the control distribution), ID exceeded Observe (paired $t(13) = 2.60$, $p = .022$) and the three distal facets ($p < .001$) and tied the most stringent, most theory-aligned comparator, Non-reactivity ($t(13) = 1.69$, $p = .12$). The defensible claim is that ID matches the best decentering-relevant facet and exceeds all others while being a single text-derived score rather than a multi-item questionnaire—a comparison that partly reflects ID’s purpose-trained, continuous format rather than its being a better measure per se (Supplementary Materials S9).

The group separation is not an artifact of practitioners writing more: word count was non-significant alongside group in an analysis of covariance ($p = .34$), and ID’s criterion associations survive verbosity control—higher ID tracked lower perceived stress ($\beta = -.34$; given the a-priori directional prediction, one-sided $p = .038$, two-sided $.076$; controlling practice group, age, and verbosity; verbosity-controlled partial $r = -.44$, $p = .004$) and higher self-compassion ($\beta = +.34$, $p = .021$; verbosity-controlled partial $r = +.46$). These are cross-sectional group differences and cannot by themselves separate a practice effect from self-selection into practice; the longitudinal analysis next speaks to within-person change.

Practice exposure over time (aging cohort). The aging cohort came from a randomized meditation trial, but the present secondary analysis had limited practice ascertainment and substantial exposure overlap: practice follow-up was available for only a subset, and 12 of 20 reached controls reported some meditation practice. The arm-level available-case contrast was null at 12 months ($g = 0.02$, 95% CI $[-0.35, 0.40]$, $p = .90$; in the compliance-measured subsample, $g = 0.27$, $p = .40$) and at 24 months ($g = -0.15$, 95% CI $[-0.63, 0.32]$, $p = .53$), with outcome availability not differing significantly across arms (Supplementary Materials S11). These estimates concern assignment in the available secondary dataset; exposure overlap would tend to attenuate an assignment effect, while missing outcome and practice data limit inference.

Analyzed instead by reported sitting-meditation exposure among those reached—a post-hoc, non-randomized analysis subject to selection on who was reached and who chose to practice—participants with any sitting practice showed greater 24-month ID change than those with none ($g = 0.80$, 95% CI $[0.13, 1.56]$, $p = .038$; Figure 4). Sitting-meditation frequency correlated with 24-month ID change (Spearman $\rho = .36$, 95% CI $[.05, .61]$, $p = .017$; 12-month $\rho = .29$, $p = .042$), but the continuous association attenuated after controlling baseline ID (partial $r = .13$, $p = .40$).

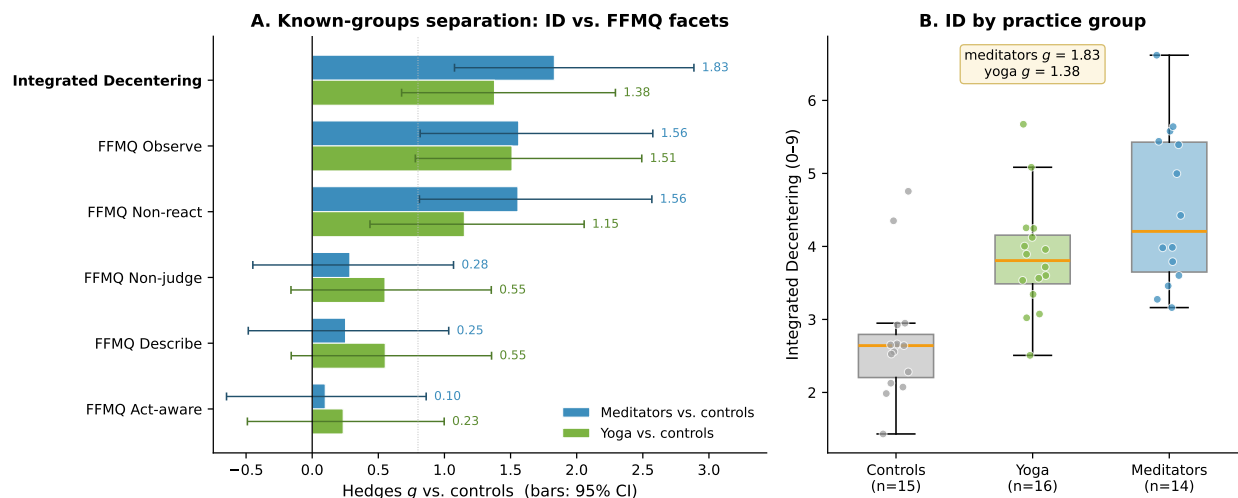


Figure 3: **Practice known-groups separation and facet specificity (neuroimaging cohort, $N = 45$).** (A) Hedges g vs. controls for ID and the five FFMQ facets, for meditation and yoga practitioners; ID exceeds every facet for meditators and matches the strongest facet (Observe) for yoga, with only the decentering-relevant facets (Observe, Non-react) discriminating strongly (error bars are 95% CIs). (B) ID by group: controls < yoga < meditators ($g = 1.83$ and 1.38 vs. controls).

The categorical contrast remained under baseline adjustment (adjusted mean difference = 0.65 ID points, $p = .009$; adjusted $d \approx 0.9$; script `28_robustness_followups.py`). In a joint model among reached participants, reported practice was associated with 24-month ID change ($p = .01$) whereas assigned arm was not ($p = .50$). Because exposure was not randomized and practice data were selectively available, these associations cannot distinguish an effect of practice from self-selection or other confounding. A sensitivity analysis coding never-reached participants as non-practicing preserved the categorical difference ($g = 0.73$), while the continuous dose association attenuated (Supplementary Materials S11).

5 Discussion

5.1 Main contribution

We introduced Integrated Decentering, a language-based measure that scores the decentering expressed in short reflective English text. Rather than treat it as an automated text score, we evaluated it as a psychological measurement instrument, asking whether it behaves like one across seven heterogeneous samples and multiple sources of validity evidence. The measure converged with both established decentering self-report scales (EQ-D $r = .36$; the smaller MPoDT sample corroborating); predicted the theoretically implicated outcomes—empathic concern, search-for-resolution in interpersonal conflict, and value-behavior coherence beyond broad personality—while remaining null on the adjacent constructs the theory does not implicate, including rumination and negative affect; and showed no positive discrimination in a length-matched impersonal argumentative register where the construct does not apply. It distinguished contemplative practitioners from matched non-practitioner controls (meditation $g = 1.83$), numerically more strongly for meditation than for yoga; in a post-hoc exposure analysis, 24-month gains tracked reported sitting practice; and aggregate scores were stable over 12 months. The contribution is the pattern, not any single effect.

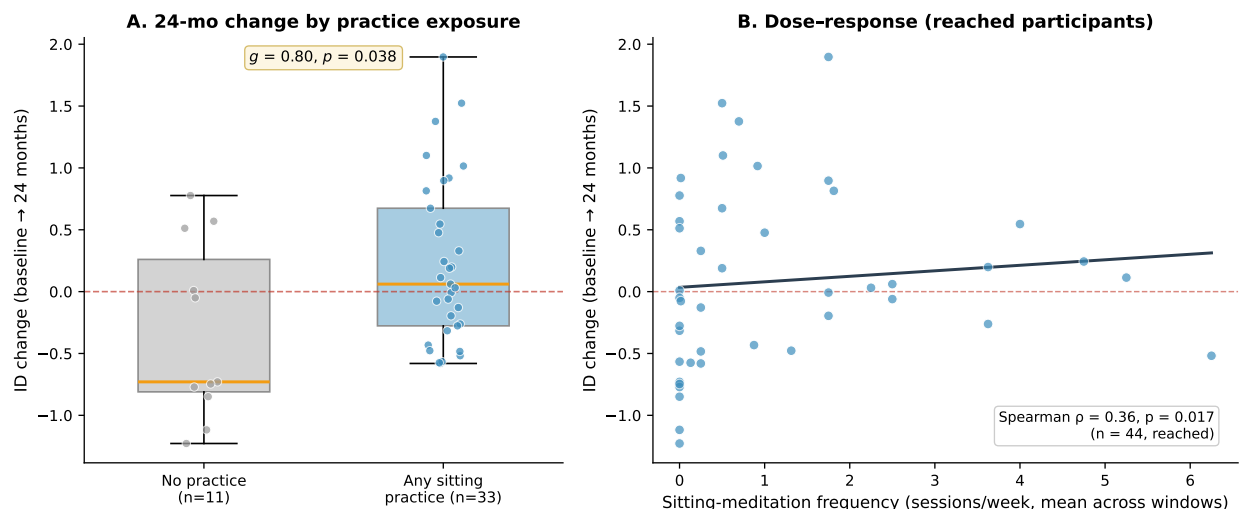


Figure 4: **Sensitivity to contemplative practice (aging cohort)**. Post-hoc associations with reported sitting-meditation exposure among participants reached for practice follow-up. (A) 24-month ID change, any sitting practice vs. none ($g = 0.80, p = .038$). (B) Dose-response: sitting-meditation frequency vs. 24-month ID change (Spearman $\rho = .36, p = .017$; unadjusted—attenuates to $r = .13$ controlling baseline ID, see text). These non-randomized exposure analyses are subject to selection and confounding.

Each tested alternative account is addressed by a different cell of the design (Table 3), and no tested alternative explains the full set of results.

A second line of support is transfer: one fixed scorer, never tuned on any validation label, carried across all seven samples and matched in-domain supervised text models where they generalized and exceeded them where they did not (three samples; Supplementary Materials S15), as expected of a measurement instrument rather than a sample-specific predictor. On the well-powered convergent test it also showed convergent signal where every tested prompted model showed none—its own untuned base models and the most capable general-purpose systems available alike (matched-wave $r = .31$ vs. $\leq .14$), across prompting with the definition alone, with scored examples, and with reasoning enabled (Supplementary Materials S15); the signal lies in the contrastive fine-tuning rather than the base model or its scale. This comparison should not be read as a general claim about the limits of such models under other prompting regimes; it shows only that prompting, at every capability level we tested, did not substitute for construct-targeted training on these data. To our knowledge this is the first text-derived decentering measure validated at this breadth, opening the language channel to the large-scale, multi-sample study that self-report decentering measures have long supported. For adopting researchers, transfer is the practical property: the released scorer can be applied to new samples of reflective text without retraining, within the scope conditions and usage guidance below.

Three design choices help define what ID is and is not. First, the measure is built to *fail* where the construct does not apply: on length-matched impersonal text carrying no self-relevant or emotional stake—the CMV persuasion corpus—it shows no signal beyond length, and we treat that boundary as positive evidence rather than a shortcoming. Second, the scorer is trained with a contrastive within-stage objective so that it tracks decentering expression rather than the coarse developmental *stage* it co-occurs with. Third, scoreability is treated as part of the construct: decentering is meaningful only in text containing something the writer could be fused with—a

feeling, belief, value, identity, or perspective—so the eligibility filters operationalize a precondition of the construct, not a preprocessing convenience.

5.2 What “integrated” means

We call the measure *Integrated* Decentering because the evidence suggests that what ID captures is not merely a momentary state but the degree to which momentary decentering has shaped a person’s habitual meaning-making—a more consolidated disposition, indexed at the per-person aggregate level. Four features support this reading: (i) *stability*: ID showed year-long rank-order stability in the active-control arm not assigned meditation, despite some self-initiated practice (test–retest $r = .69$), in the range of established trait measures; (ii) *behavioral consequence*: it predicts value–behavior coherence beyond broad personality; (iii) *practice-linked change*: it is associated with sustained meditation exposure; and (iv) *developmental overlap without reduction*: it shares a maturity-relevant axis with a human expert’s developmental score ($r = .79$, as theory predicts) yet shows decentering-specific predictive structure beyond it, including decentering-relevant self-report net of that score. A measure that is stable, behaviorally consequential, sensitive to practice exposure, and maturity-relevant-but-not-reducible behaves like an index of a consolidated capacity rather than a transient state. This supports a trait-like interpretation for aggregates, while the score itself remains a property of text.

5.3 Relevance to mindfulness science

Contemplative research has long faced a gap between practice delivered and decentering internalized, with practice hours and trait-mindfulness self-report each capturing only part of it. ID offers a complementary, low-cost outcome. Two findings are relevant here. First, the known-groups pattern was specific: ID distinguished practitioners from controls (more so for seated meditation than for yoga), and within meditators it tied the most stringent decentering-relevant facet (Non-reactivity) while exceeding the rest. Second, the aging cohort showed a null assignment contrast alongside an association between reported practice and 24-month ID change among those with practice follow-up. Because that exposure analysis was post-hoc, non-randomized, and restricted to a selected subset, it is hypothesis-generating and does not establish that practice caused the change. It motivates preregistered testing of ID as a process or mediator measure in a trial that preserves assignment-to-exposure separation. We do not propose ID as a replacement for self-report or practice logs—each answers a question ID cannot—but as a third channel for triangulation.

5.4 Relevance beyond meditation

Decentering is a mechanism of broad reach, and a portable, scalable measure of it is useful wherever the capacity is a target but outcome infrastructure lags. The criterion evidence here already spans several such domains: empathic concern in response to others’ suffering (WASSA), wise reasoning about interpersonal conflict (Grossmann), and the alignment of values with action (MTurk). The same instrument is therefore relevant to psychotherapy process research—decentering is the proposed common mechanism of MBCT and the cognitive defusion of ACT—and to work on emotion regulation, conflict and negotiation, empathy, and wisdom, where reflective distance from one’s own standpoint is implicated but is hard to measure at scale. Where those traditions measure the construct by self-report—for ACT, the Cognitive Fusion Questionnaire [Gillanders et al., 2014]—ID offers a complementary text-based index that can be applied to reflective writing those literatures already collect, though it has not yet been validated against a fusion-specific scale.

A further direction concerns reflective agency in human–AI interaction: as conversational systems increasingly shape how people reason, a text-derived measure of whether someone holds an introduced interpretive frame as contestable—rather than operating through it as their own—may help index the receiver-side reflective capacity that such interactions can support or erode. We flag this as a forward direction rather than a present claim: conversational dialogue turns differ in register from the reflective writing validated here, so applications to conversational agents or digital mental health settings would require in-domain validation first; the validation here is of a decentering measure, not of any application.

5.5 Developmental maturity: benchmark and boundary condition

Developmental maturity is the closest benchmark for ID and a boundary condition on its neural correlates. ID correlates with an expert Total Weighted Score (TWS) at $r = .79$. This overlap is expected—a shared maturity-relevant axis that ego development and decentering are theorized to hold in common [Singleton et al., 2021], and same-protocol scoring—so we do not treat the correlation’s size as independent evidence for ID; equally, ID was not trained as a stage regressor (Section 2). Distinctness rests on differential prediction, not on a low correlation: constructs accepted as distinct routinely correlate in this range [Cronbach and Meehl, 1955, Campbell and Fiske, 1959]. Against TWS, ID tracked self-compassion more closely and added unique variance beyond it for both self-compassion and perceived stress, while diverging in sign on quality of life (Section 4.5); about 62% of ID’s variance is shared with TWS and roughly 38% is not. Within the maturity-related space, ID indexes a specific, decentering-relevant capacity rather than overall developmental level.

As a boundary condition, the exploratory neuroimaging analysis is worth reporting in brief. Computed from text alone, ID re-derived (rather than independently discovered) a developmentally gated pattern of posterior-cingulate connectivity previously reported for this cohort [Singleton et al., 2021], including a whole-sample PCC–IFC correlate ($r \approx .54$; see S13 for n , confidence interval, and exact p) that is itself developmentally shared rather than decentering-specific. The most suggestive, and at the same time weakest, result is a stage-dependent sign inflection in the PCC–TPJ and PCC–dmPFC edges, negative below the transition to self-authoring and positive above, with the most theoretically decentering-aligned edge, PCC–TPJ, plausibly indexing the coupling of self-referential processing to perspective-taking and self–other distinction.

The caveats are heavy: it rests on an underpowered $n = 45$ split, an ID×stage interaction significant only at $p = .046$ uncorrected that does not survive multiple-comparison correction over edges and stages, and a legacy 5-minute 1.5T acquisition with inherited preprocessing. It is also a non-independent reanalysis of the very cohort whose published pattern it reproduces, in which ID overlaps the expert developmental rating at $r = .79$, so that the gating is developmental and shared with that rating rather than decentering-unique. Stage also remains confounded with contemplative practice. So bounded, the result is best read as one automated, language-derived score tracking a previously expert-gated brain pattern in the same cohort (full statistics and figure in S13), a corroborating extension, not a load-bearing pillar of the construct claim.

5.6 Alternative interpretations

The interpretation of ID survives the alternative accounts the data can test: each is addressed by a different part of the design, so no single confound accounts for the whole pattern (Table 3). The account most closely entwined with ID—developmental maturity—was addressed in the preceding subsection: ID is best read as a maturity-*relevant* but not maturity-*reducible* index of decentering.

Table 3: Alternative accounts and the diagnostic tests that bear on each. Each is addressed by a different part of the design; no single account explains the whole pattern.

Account	Diagnostic test	Result	Verdict
Length / verbosity	Length controls; length-matched CMV boundary	No positive discrimination after matching (AUC .45 [.40, .49])	Not verbosity; do not apply to impersonal, non-self-relevant content
Verbal ability	AMNART-EVIQ adjustment	ID β shifts $\leq .02$; AMNART n.s.	Not verbal IQ
Broad personality	Big Five incremental (cross-validated)	VBC $\Delta R^2 = .027$ (CV .022)	Not reducible to personality
Generic positivity / good reflection	Wise-reasoning subscale specificity; rumination / neg-affect nulls	Search-for-resolution +; adjacent subscales and rumination null	Selective, not generic “good reflection”
Surface linguistic distancing	LIWC-22 distancing index; incremental ΔR^2	Distancing ≈ 0 with criterion in both large samples; ID adds $\Delta R^2 = .13$ (WASSA), .04 (Grossmann); distancing adds $\leq .01$	Not a relabeling of surface markers
Demographics	Age, sex, education adjustment	ID $\times$ demographics $\leq .13$; effects survive	Not a demographic proxy
Developmental maturity	Expert TWS benchmark; anti-shortcut training; differential-prediction tests	$r = .79$ theory-expected overlap; tracks self-compassion more closely, diverges in sign for quality of life, predicts decentering-relevant self-report net of TWS	Shared axis + same protocol; decentering-specific, not reducible
Generic LLM scoring	Zero-/few-shot prompting; base-model ablation	ID shows convergent signal where base and prompted general-purpose models do not (zero-shot, few-shot, and reasoning) on the better-powered convergent test	Signal from the trained scorer, not just prompting

5.7 Limitations and future directions

Several limitations bound these conclusions. *Language and register*: all validation was in English, and the scorer is validated for self-relevant, reflective or emotionally engaged text; on impersonal content with no such stake (e.g. persuasive argument about an external topic, or declarative statements) it can reduce to a length proxy, as the single length-matched CMV test indicates. *Aggregation*: the validity reported here is established at the aggregate level (per person, or per text or event for the essay and diary samples); single-stem scores should not be read on their own.

Scorer dependence: the specific number a passage receives depends on the model. ID has not yet been benchmarked against blinded human ratings of decentering *expression* in the same passages; convergent validity here rests on two validated self-report decentering scales (EQ-D, MPoDT), not on direct text-level human agreement (ID does predict trained-rater codings of text on the adjacent wise-reasoning constructs, but no rater judged decentering expression itself). Pending that benchmark, ID is best applied to within-study contrasts—group differences and change over time—rather than to absolute, cross-sample score levels. Scoring stability across prompt paraphrases and measurement invariance across the heterogeneous genres also remain to be established. *Sample and design*: the neuroimaging cohort was small ($n = 45$), and the aging analyses used available secondary data with incomplete long-term outcomes and practice ascertainment. ID was not a prespecified trial outcome; all aging-cohort ID analyses were post-hoc and not preregistered. The

reported-practice analyses were non-randomized and restricted to participants with follow-up, so they are vulnerable to selection, self-report error, and unmeasured confounding and should not be interpreted causally. The neural analysis is likewise a non-independent reanalysis of the cohort whose published pattern it is consistent with.

Construct and elicitation coverage. Within the three-process model of decentering [Bernstein et al., 2015], ID aligned most clearly with the components most directly sampled by reflective sentence completions: meta-awareness ($r = .49$) and disidentification ($r = .43$). Its association with reduced reactivity was weaker ($r = .17$). Text can express reactivity and its absence, but these sentence completions provide brief, dispositional reflections rather than in-the-moment affective challenges. Reduced reactivity may therefore be undersampled here. The estimate is imprecise ($n = 35$; approximate 95% CI $[-.17, .48]$), so this is a hypothesis about elicitation and construct coverage rather than a settled boundary. Future validation should test ID against state-level, behavioral, linguistic, and physiological indices of reduced reactivity. A related note concerns the scale’s low end: a low aggregate indicates little *expressed* decentering, not necessarily active fusion or distress—consistent with the rumination and negative-affect nulls—and discrimination among strongly fused registers is anchored here only by the role-play coping gradient (Supplementary Materials S12). Applications that specifically target fusion should validate that pole directly.

Guidance for research use. Practical recommendations for studies adopting ID—on matching scored-text quantity to the reliability required, register eligibility, per-aggregate reporting of passage and word counts, verbosity control, and filter-view reporting—are set out in Supplementary Materials S1.1.

The corresponding agenda is clear. First and most important, scorer-independent validation against human judgement: blinded human ratings of decentering *expression* in the same passages across multiple genres, with inter-rater reliability and formal human–model agreement—the most direct check of the construct, and the top priority before operational use. Alongside it, robustness to instructed or coached “decentered-style” writing remains untested: all text analyzed here was written without knowledge of the scorer, so score inflation under deliberate impression management is an open question for elicited-text applications. Second, pre-registered replication of the practice sensitivity in a design that preserves the assignment-to-exposure contrast (e.g., a higher-barrier intervention, or contamination-monitored trials), and of the criterion associations in fresh samples. Third, independent neuroimaging with modern acquisition to test the PCC–TPJ association outside the originating cohort. Fourth, cross-cultural and non-English adaptation. Finally, application studies in the adjacent domains above—psychotherapy process, emotion regulation, conflict, and reflective agency in human–AI dialogue—where a scalable decentering measure could fill a standing gap.

5.8 Conclusion

Integrated Decentering is a brief, automated language measure of decentering expressed in reflective text. Across convergent, nomological, behavioral, discriminant, and practice-linked tests in seven heterogeneous samples, the pattern matched what the construct requires—with informative nulls and a clear boundary—and held beyond text length, verbal ability, and personality; with respect to developmental maturity, ID shares a maturity-relevant axis with a human expert’s score but is not reducible to it. ID can sit alongside self-report and practice-exposure measurement as a scalable internalization index for contemplative science and the many adjacent fields where decentering is a

mechanism but measurement has lagged, pending the cross-scorer, cross-cultural, and pre-registered replication the evidence now motivates.

Data and Code Availability

Several validation datasets are public: the WASSA empathic-reactions essays [Buechel et al., 2018], the ChangeMyView “winning arguments” corpus [Tan et al., 2016], and the MTurk community sample [Fournier et al., 2022] (<https://osf.io/7adtj/>). The Grossmann conflict-diary data are available from the original authors [Grossmann et al., 2021]. The neuroimaging cohort is a secondary analysis of [Singleton et al., 2021] and the longitudinal aging cohort a post-hoc secondary analysis of a randomized trial at the Athinoula A. Martinos Center for Biomedical Imaging [Sevinc et al., 2021] (ClinicalTrials.gov NCT02628548; registered 11 December 2015). The de-identified participant-level analytic data underlying the reported aging-cohort results, together with variable descriptions, are available from the corresponding author on reasonable request following publication, for research purposes compatible with participant consent and applicable institutional privacy requirements, and for as long as the data are retained; requests should identify the proposed analyses. The parent trial’s methods and CONSORT flow diagram are publicly available with its primary report. De-identified neuroimaging-cohort data can be requested from the authors. The MPoDT convergent sample was collected by the authors via Prolific; de-identified data are available from the authors on request. The reflective sentence-completion training material draws on WUSCT-tradition and adult-developmental corpora [Lanning et al., 2018, Madl and Cook-Greuter, 2026] that retain their own terms and are not redistributed; inquiries should be directed to the authors.

The Integrated Decentering scorer (a two-adaptor Qwen LoRA ensemble) is released as a gated open-weight model (access granted on request for non-commercial research use) with runnable inference and eligibility-filtering code under a CC-BY-NC-4.0 license at <https://huggingface.co/tmadl/ID-decentering-scorer-ensemble>. The released weights are fixed and versioned, so a result reported against ID refers to a specific, citable model version. A reproducibility package accompanies this paper containing the complete, ordered analysis code together with the reference outputs and figures it produces, so that every number, table, and figure reported here can be traced to the exact script that generated it; the build step verifies the released scorer’s documented validation correlations (MPoDT, EQ-D) exactly, confirming the pipeline. Consistent with the scorer’s gated, non-commercial, research-only release, the per-passage ID scores the scripts consume are available to researchers on request rather than posted publicly, as is the text-free scored intermediate for the supplementary ECSC/COPING check (Section S12; the underlying role-play corpus [Troiano et al., 2024] is publicly available from its authors and is not redistributed here). Analyses used Python 3.10 (NumPy, SciPy, pandas, statsmodels). The full reproducibility code (scripts, reference outputs, and figures) is available on the Open Science Framework at https://osf.io/ay532/?view_only=8250955e50124d60b3b095956a63b540.

Declarations

Funding. No funding was received for this work. The present study is a secondary analysis of previously collected data; funding for the original data collections is reported in their respective source publications [Gard et al., 2014, Singleton et al., 2021, Sevinc et al., 2021, Fournier et al., 2022] (the longitudinal aging trial was supported by National Institute on Aging award R01 AG048351; the neuroimaging cohort by NIH award R21 AT003673).

Competing interests. T.M. developed and openly released the Integrated Decentering scoring model for non-commercial research use (CC-BY-NC-4.0). T.M.’s authorship of the released instrument constitutes a potential competing interest, which we mitigate as follows: (a) the model parameters were fixed and versioned before any validation sample was scored; (b) all validity tests—including the predicted nulls and the predicted discriminant failure—are reported regardless of outcome; (c) a complete reproducibility package regenerates every reported number; (d) the scorer is available at no cost for non-commercial research, permitting independent—including adversarial—validation, which we explicitly invite; and (e) S.W.L., who has no stake in the instrument, had full access to the cohort data and reviewed all reported analyses. The gating on the released weights exists to prevent misuse for evaluating identifiable individuals, not to restrict scientific scrutiny, which we invite; the scorer is explicitly not validated or licensed for decisions about identifiable individuals. S.W.L. declares no competing interests.

Ethics approval. The neuroimaging study and the longitudinal aging trial were approved by the Partners Healthcare Institutional Review Board (Massachusetts General Hospital) [Singleton et al., 2021, Sevinc et al., 2021]. The MTurk study was approved by the Institutional Review Board at Florida Atlantic University [Fournier et al., 2022]. The public NLP corpora (WASSA, ChangeMyView) were analyzed under their respective data-use terms. The MPoDT study was reviewed by the OFAI-REB (OF AI Review Board) at the Austrian Research Institute for Artificial Intelligence (OF AI). The Board determined that the study was exempt from full ethics review because it comprised anonymous, minimal-risk, non-medical behavioral survey research falling outside the statutory remit for mandatory medical-research ethics review; institutional ethics screening and informed-consent safeguards were nevertheless retained (determination/reference no. OF AI-REB-2026-001). All participants provided informed consent before participation. The study was conducted in accordance with the Declaration of Helsinki and applicable Austrian requirements. The other primary studies were approved by the ethics bodies named above and likewise conducted in accordance with the Declaration of Helsinki.

Consent to participate. All participants in the primary studies provided written informed consent; participants in the MPoDT online sample provided informed consent prior to participation. The public NLP corpora were collected and released by third parties under their own consent procedures.

Consent for publication. No identifiable individual data are reported; consent for publication is not applicable.

Use of artificial intelligence. The instrument evaluated in this article is itself a language-model-based scorer, described in Section 2 and the Supplementary Materials. Beyond that, the authors used generative AI tools to assist with coding and language editing; the authors carried out all analyses and writing, verified every reported number against the released reproducibility code, and take full responsibility for the content.

Authors’ contributions. T.M. conceived the study, developed and released the scoring system, performed all statistical analyses, and wrote the manuscript. S.W.L. provided access to the neuroimaging and aging datasets, contributed to interpretation of the neural findings, and critically revised the manuscript. Both authors approved the final version.

References

- [Bernstein et al., 2015] Bernstein, A., Hadash, Y., Lichtash, Y., Tanay, G., Shepherd, K., and Fresco, D. M. (2015). Decentering and related constructs: A critical review and metacognitive processes model. *Perspectives on Psychological Science*, 10(5):599–617.
- [Bowles et al., 2022] Bowles, N. I., Davies, J. N., and Van Dam, N. T. (2022). Dose–response relationship of reported lifetime meditation practice with mental health and wellbeing: A cross-sectional study. *Mindfulness*, 13(10):2529–2546.
- [Boyd et al., 2022] Boyd, R. L., Ashokkumar, A., Seraj, S., and Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. Technical report, University of Texas at Austin, Austin, TX.
- [Bradley and Terry, 1952] Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- [Brienza et al., 2018] Brienza, J. P., Kung, F. Y. H., Santos, H. C., Bobocel, D. R., and Grossmann, I. (2018). Wisdom, bias, and balance: Toward a process-sensitive measurement of wisdom-related cognition. *Journal of Personality and Social Psychology*, 115(6):1093–1126.
- [Buechel et al., 2018] Buechel, S., Buffone, A., Slaff, B., Ungar, L., and Sedoc, J. (2018). Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4758–4765.
- [Campbell and Fiske, 1959] Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105.
- [Connelly and Ones, 2010] Connelly, B. S. and Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers’ accuracy and predictive validity. *Psychological Bulletin*, 136(6):1092–1122.
- [Cook-Greuter, 1999] Cook-Greuter, S. R. (1999). *Postautonomous Ego Development: A Study of Its Nature and Measurement*. Ed.d. dissertation, Harvard University Graduate School of Education. Abstract published in Dissertation Abstracts International, 60(6-B), 3000.
- [Cronbach and Meehl, 1955] Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.
- [Fournier et al., 2022] Fournier, M. A., Dong, M., Quitasol, M. N., Weststrate, N. M., and Di Domenico, S. I. (2022). Components and correlates of personality coherence in action, agency, and authorship. *European Journal of Personality*, 36(3):413–434.
- [Fresco et al., 2007] Fresco, D. M., Moore, M. T., van Dulmen, M. H. M., Segal, Z. V., Ma, S. H., Teasdale, J. D., and Williams, J. M. G. (2007). Initial psychometric properties of the Experiences Questionnaire: Validation of a self-report measure of decentering. *Behavior Therapy*, 38(3):234–246.
- [Gard et al., 2014] Gard, T., Taquet, M., Dixit, R., Hölzel, B. K., de Montjoye, Y.-A., Brach, N., Salat, D., Dickerson, B. C., Gray, J. R., and Lazar, S. W. (2014). Fluid intelligence and brain functional organization in aging yoga and meditation practitioners. *Frontiers in Aging Neuroscience*, 6:76. PMID: 24795629.

- [Gard et al., 2015] Gard, T., Taquet, M., Dixit, R., Hölzel, B. K., Dickerson, B. C., and Lazar, S. W. (2015). Greater widespread functional connectivity of the caudate in older adults who practice kripalu yoga and vipassana meditation than in controls. *Frontiers in Human Neuroscience*, 9:137. PMID: 25852521.
- [Gignac and Szodorai, 2016] Gignac, G. E. and Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102:74–78.
- [Gillanders et al., 2014] Gillanders, D. T., Bolderston, H., Bond, F. W., Dempster, M., Flaxman, P. E., et al. (2014). The development and initial validation of the cognitive fusion questionnaire. *Behavior Therapy*, 45(1):83–101.
- [Grossman, 2008] Grossman, P. (2008). On measuring mindfulness in psychosomatic and psychological research. *Journal of Psychosomatic Research*, 64(4):405–408.
- [Grossmann et al., 2021] Grossmann, I., Dorfman, A., Oakes, H., Santos, H. C., Vohs, K. D., and Scholer, A. A. (2021). Training for wisdom: The distanced-self-reflection diary method. *Psychological Science*, 32(3):381–394.
- [Gu et al., 2015] Gu, J., Strauss, C., Bond, R., and Cavanagh, K. (2015). How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and well-being? A systematic review and meta-analysis of mediation studies. *Clinical Psychology Review*, 37:1–12.
- [Hanley et al., 2020] Hanley, A. W., Bernstein, A., Nakamura, Y., Hadash, Y., Rojas, J., Tennant, K. E., Jensen, R. L., and Garland, E. L. (2020). The Metacognitive Processes of Decentering Scale: Development and initial validation of trait and state versions. *Psychological Assessment*, 32(10):956–971.
- [Kegan, 1982] Kegan, R. (1982). *The Evolving Self: Problem and Process in Human Development*. Harvard University Press, Cambridge, MA.
- [Kross and Ayduk, 2017] Kross, E. and Ayduk, O. (2017). Self-distancing: Theory, research, and current directions. In *Advances in Experimental Social Psychology*, volume 55, pages 81–136. Academic Press.
- [Lanning et al., 2018] Lanning, K., Pauletti, R. E., King, L. A., and McAdams, D. P. (2018). Personality development through natural language. *Nature human behaviour*, 2(5):327–334.
- [Lau et al., 2006] Lau, M. A., Bishop, S. R., Segal, Z. V., Buis, T., Anderson, N. D., Carlson, L., Shapiro, S., Carmody, J., Abbey, S., and Devins, G. (2006). The Toronto Mindfulness Scale: Development and validation. *Journal of Clinical Psychology*, 62(12):1445–1467.
- [Lebois et al., 2015] Lebois, L. A. M., Papies, E. K., Gopinath, K., Cabanban, R., Quigley, K. S., Krishnamurthy, V., Barrett, L. F., and Barsalou, L. W. (2015). A shift in perspective: Decentering through mindful attention to imagined stressful events. *Neuropsychologia*, 75:505–524.
- [Madl and Cook-Greuter, 2026] Madl, T. and Cook-Greuter, S. R. (2026). Evolutionary ai-driven discovery of the psychological structure of adult ego development. Manuscript submitted for publication.

- [Nook et al., 2017] Nook, E. C., Schleider, J. L., and Somerville, L. H. (2017). A linguistic signature of psychological distancing in emotion regulation. *Journal of Experimental Psychology: General*, 146(3):337–346.
- [Parsons et al., 2017] Parsons, C. E., Crane, C., Parsons, L. J., Fjorback, L. O., and Kuyken, W. (2017). Home practice in mindfulness-based cognitive therapy and mindfulness-based stress reduction: A systematic review and meta-analysis of participants’ mindfulness practice and its association with outcomes. *Behaviour Research and Therapy*, 95:29–41.
- [Pennebaker et al., 2014] Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLoS ONE*, 9(12):e115844.
- [Sevinc et al., 2021] Sevinc, G., Rusche, J., Wong, B., Datta, T., Kaufman, R., Gutz, S. E., Schneider, M., Todorova, N., Gaser, C., Thomalla, G., Rentz, D., Dickerson, B. D., and Lazar, S. W. (2021). Mindfulness training improves cognition and strengthens intrinsic connectivity between the hippocampus and posteromedial cortex in healthy older adults. *Frontiers in Aging Neuroscience*, 13:702796.
- [Shapiro et al., 2006] Shapiro, S. L., Carlson, L. E., Astin, J. A., and Freedman, B. (2006). Mechanisms of mindfulness. *Journal of Clinical Psychology*, 62(3):373–386.
- [Singleton et al., 2021] Singleton, O., Newlon, M., Fossas, A., Sharma, B., Cook-Greuter, S. R., and Lazar, S. W. (2021). Brain structure and functional connectivity correlate with psychosocial development in contemplative practitioners and controls. *Brain Sciences*, 11(6):728.
- [Steiger, 1980] Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.
- [Strohmaier, 2020] Strohmaier, S. (2020). The relationship between doses of mindfulness-based programs and depression, anxiety, stress, and mindfulness: A dose-response meta-regression of randomized controlled trials. *Mindfulness*, 11:1315–1335.
- [Tan et al., 2016] Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 613–624.
- [Teasdale et al., 2002] Teasdale, J. D., Moore, R. G., Hayhurst, H., Pope, M., Williams, S., and Segal, Z. V. (2002). Metacognitive awareness and prevention of relapse in depression: Empirical evidence. *Journal of Consulting and Clinical Psychology*, 70(2):275–287.
- [Troiano et al., 2024] Troiano, E., Labat, S., Stranisci, M. A., Damiano, R., Patti, V., and Klinger, R. (2024). Dealing with controversy: An emotion and coping strategy corpus based on role playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1634–1658, Miami, Florida, USA. Association for Computational Linguistics.
- [Van Dam et al., 2018] Van Dam, N. T., van Vugt, M. K., Vago, D. R., Schmalzl, L., Saron, C. D., Kerr, C. E., Lazar, S. W., et al. (2018). Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on Psychological Science*, 13(1):36–61.

[van der Velden et al., 2015] van der Velden, A. M., Kuyken, W., Wattar, U., Crane, C., Pallesen, K. J., Dahlgaard, J., Fjorback, L. O., and Piet, J. (2015). A systematic review of mechanisms of change in mindfulness-based cognitive therapy in the treatment of recurrent major depressive disorder. *Clinical Psychology Review*, 37:26–39.

Supplementary Materials

Measuring Decentering in Reflective Language: Construct Validation of a Text-Based Scorer

All numbers below are regenerated by the accompanying reproducibility package from the per-text scorer outputs; the build step also verifies the released scorer’s documented validation correlations (MPoDT, EQ-D) exactly, confirming the pipeline.

S1 Scorer architecture and training

The released scorer is an equal-weight ensemble of two independently trained LoRA adapters (Table S1). Each member scores text by *logit-expected-value decoding*: a single forward pass over the assistant scoring position reads the logits at the ten digit tokens “0”–“9”, applies a softmax, and returns $\sum_{d=0}^9 d \cdot \text{softmax}(\text{logits})[d] \in [0, 9]$. The members are strongly correlated on training-format text ($r \approx .93$) but not interchangeable across the validation battery, which motivates equal-weight averaging rather than selecting a single member or fitting validation-specific weights. Single-stem reliability is low by design— $\text{ICC}(1, 1) \approx .05\text{--}.09$, since one short completion carries little stable signal—which is why these single-sentence cohorts are aggregated to the per-person or per-condition level, where reliability is adequate ($\text{ICC}(1, k) = .75$ in the neuroimaging cohort, $.65$ in the MPoDT sample, and $.66$ per person-wave in the aging cohort). Because signal accrues with the quantity of scored text, reaching person-level reliability $\approx .70$ takes roughly 25–44 single sentence completions (Spearman–Brown from the single-stem ICC; `15_reliability.py`); longer first-person passages such as diary entries or essays carry more text per response and are expected to need correspondingly fewer.

Table S1: Ensemble members.

	v9_best	v13_qwen36_best
Base model	Qwen3.5-27B (4-bit QLoRA)	Qwen3.6-27B (4-bit QLoRA)
LoRA	$r=32, \alpha=64$	$r=16, \alpha=32$
Init	warm-start from prior SCT LoRA	fresh
Training pairs	$\approx 62\text{k}$	$\approx 106\text{k}$ (broader content)
Objective	Bradley–Terry pairwise loss on $E[\text{digit softmax}]$	

Contrastive target and anti-shortcut training. Rather than regressing onto a developmental-stage label, members are trained on length-matched pairs to predict which passage expresses more decentering. The proxy target is a decentering-aligned indicator from upstream adult-developmental work [Madl and Cook-Greuter, 2026], treated as deliberately imperfect. Pairs are of three kinds, weighted to contribute approximately equally to the loss (Table S2). The *within-stage* pairs—same developmental stage, different decentering proxy—are the anti-shortcut signal: they force the model to track reflective distance even when stage is held constant, discouraging stage/bucket shortcut learning. They are the smallest bucket and are correspondingly up-weighted.

Training material comprised reflective sentence-completion corpora in the Loevinger/WUSCT tradition with adult-developmental extensions [Lanning et al., 2018, Madl and Cook-Greuter, 2026]. The scorer is English-only, validated for reflective text of roughly sentence-to-paragraph length, and intended for aggregate (per-person/per-condition) use, not for decisions about identifiable individuals.

Table S2: Bradley–Terry pair types and loss weights.

Pair type	Role	Weight
cross_bucket	main high-vs-low decentering contrast	0.44
within_pc	graded contrast within later adult-developmental material	1.64
within_stage	same-stage contrast (anti-shortcut)	9.43

S1.1 Scorer datasheet

We summarize the released scorer in datasheet form; architecture, training pairs, and reliability are in Section S1, and full training and inference code with versioned weights accompany the public model card.

Training target (what it is). Each member was trained to rank pairs of passages by expressed decentering using a single *decentering-aligned indicator* drawn from the upstream adult-developmental work [Madl and Cook-Greuter, 2026] as a deliberately imperfect, theory-motivated proxy. The indicator is one signal selected for its alignment with stepping back from and holding one’s thoughts, affects, and identities as transient mental events; it orders training pairs by relative decentering only—it is not a regression target and not a developmental-stage or ego-level label.

What it is not. The target is not a Loevinger/WUSCT stage code, not a Total Weighted Score, and not any single Loevinger sentence-completion category; the adult-developmental corpora supply reflective text and contrastive structure, not the definition of the construct. The up-weighted within-stage pairs (Table S2) hold developmental stage constant, pushing the scorer to track reflective distance rather than maturity level; the residual overlap with an expert developmental score is reported and bounded (Sections S9, S14), and discriminant specificity rests on the large non-developmental criterion corpora (WASSA, Grossmann, MTurk), not on the small developmental cohort.

Inputs. English, first-person or self-relevant reflective text of roughly sentence-to-paragraph length, containing something the writer could be fused with (a belief, value, feeling, identity, or perspective). Purely factual or impersonal text is out of scope (main text, *Discriminant validity and construct boundaries*).

Output and score distribution. A continuous 0–9 score per passage (logit-expected-value decoding; Section S1). Across validation corpora the per-text score uses the full range, with genre-dependent location (Table S3); because absolute level varies by genre and elicitation, ID is intended for within-study contrasts (group differences, change over time), not for comparing absolute scores across samples (main text, *Measurement behavior and reliability*).

Intended use and out of scope. A research instrument for aggregate-level decentering expression. It is not validated for, and should not be used for, decisions about identifiable individuals, or to infer a person’s mindfulness, wisdom, mental health, or developmental stage (main text, *The Integrated Decentering scorer*).

Guidance for research use. For studies that adopt ID we recommend: (i) matching the quantity of scored text to the reliability required—single short sentence completions are individually weak estimates of the person ($\text{ICC}(1,1) \approx .05\text{--}.09$) and were aggregated over many stems per person here, while longer passages carry more signal per response—and reporting the reliability achieved; (ii) applying it only to self-relevant, reflective or emotionally engaged English text, not impersonal content lacking such a stake (e.g. declarative or external-topic writing); (iii) reporting the number of passages and the word counts that contribute to each aggregate; (iv) reporting which eligibility-filter view was used and confirming that conclusions hold across views; and (v) including a verbosity control (e.g., log word count) in any criterion analysis.

Table S3: Per-text ID score distribution by corpus (released v9v13 ensemble, 0–9 scale, unfiltered).

Corpus (text unit)	n	mean	SD	min	max
WASSA (essay)	1,860	3.42	2.75	0.0	9.0
Grossmann (event)	424	1.33	1.55	0.0	8.9
MPoDT (sentence)	629	3.74	3.41	0.0	9.0
MGH neuro (sentence)	1,552	3.63	3.40	0.0	9.0
ECSC role-play (reply)	1,200	2.81	2.91	0.0	9.0

S2 Eligibility filtering

Two pre-specifiable filters operationalize the precondition that a passage contain self-relevant material. The *mechanical* filter keeps a sentence if it contains a first-person pronoun, a belief/interior-state marker, or moderate affect ($|\text{VADER compound}| \geq 0.4$); it is deterministic and local. The *LLM v2.2* filter prompts a separate annotator model to extract spans where the precondition holds (higher recall, including author-owned external-topic stances). Filter views are subset operations on the same per-text scores. For the MGH-45 stems, per-stem retention was 77.3% (mechanical), 81.0% (v2.2), and 66.6% (their conjunction). Filtering is most consequential for mixed-register text (WASSA, where the both-filters view is primary and the unfiltered view a sensitivity check), but it also removes unscorable completions—which carry only length—from short sentence-completion data, materially de-confounding small-sample convergent estimates (Section S4).

Filter handling across analyses. We verified that no conclusion depends on the filtering choice. WASSA uses the both-filters view as primary; the MPoDT convergent estimate is reported on the scoreable subset (Section S4); MGH-45 known-groups, criterion, and ID–TWS results are stable across all/mech/v2.2/AND views (Section S14; neural in Section S13) and the group separation is not length-driven (word count $p = .34$); EQ-D is stable across views ($r = .35\text{--}.36$, strongest on the primary scoreable view); and MTurk value–behavior coherence is essentially unchanged across filter views (partial $r = .32\text{--}.34$). The Grossmann conflict-diary events are long first-person narratives that are 100% scoreable under both filters, so filtering is a no-op there ($\beta = .176$ unchanged). The MPoDT sample was the one materially affected—small n , short stems, and a non-trivial unscorable fraction—which is why its convergent estimate is reported on the scoreable subset.

S3 Sample characteristics and demographic confounds

Demographic confounds. ID is not a demographic proxy. It correlated only modestly with age ($r = .13$ MTurk, $-.15$ neuroimaging) and sex ($r \approx .10\text{--}.13$), and negligibly with education ($r = .01$,

Table S4: Validation sample characteristics. Public NLP corpora carry their own (source-documented) demographics.

Sample	N	Age	% female
MPoDT (Prolific)	35	19–59	53
MGH aging (EQ-D / practice)	143	65–80	52
MTurk (Fournier)	446	$M=48.4$ ($SD=7.3$)	53
MGH neuroimaging	45	$M=51.5$ ($SD=9.9$)	59
WASSA essays	1,860	source-documented	—
Grossmann diaries	424 / 298	source-documented	—
ChangeMyView	4,311 pairs	n/a (online text)	—

MTurk). Criterion effects survived demographic adjustment: adding sex to the neuroimaging models left the self-compassion ($\beta_{ID} = +.34$) and perceived-stress ($\beta_{ID} = -.35$) coefficients unchanged (sex non-significant), and adding age, sex, and education to the MTurk value-behavior-coherence model left the partial correlation essentially intact ($.33 \rightarrow .32$, $n = 417$). Education was unavailable for the neuroimaging and aging cohorts.

S4 Convergent validity (full)

Table S5: MPoDT sample, per-participant Pearson r ($n = 35$); EQ-D decentering, grand-average across waves.

MPoDT outcome	all stems (629)	mech-eligible (504)
Total	+.392	+.451
Meta-awareness	+.415	+.493
Disidentification	+.419	+.432
Reduced Reactivity	+.113	+.172
Composite	+.414	+.472
<i>EQ-D decentering</i> ($n = 124$):		
all stems	+.356	
mech-eligible	+.352	
mech $\wedge$ LLM-scoreable	+.362	
baseline only, mech ($n = 52$)	+.417	

Verbosity robustness. The convergent associations are length-robust once unscoreable completions (which carry only length) are excluded. Both convergent anchors are reported under the same eligibility views. Controlling completion length, EQ-D is stable: partial $r = .30$ (all) $\rightarrow .31$ (mech) $\rightarrow .32$ (v2.2) $\rightarrow .33$ (both), reflecting its low unscoreable fraction. For the MPoDT sample, the length-partialled association instead rises as the filters remove unscoreable completions: partial $r = .17$ (all stems, ns) $\rightarrow .28$ (mechanical) $\rightarrow .27$ (LLM v2.2) $\rightarrow .32$ (both, $p = .06$), while the ID-length coupling falls from $r = .67$ to $.57$. Stripping the fixed stems (completion-only vs. full-sentence length) makes no material difference, because the stem set is common to all participants. Known-groups separation and the criterion associations are also length-robust (word count was non-significant alongside group, $p = .34$; SCS and PSS length-controlled partials $+.46$ and $-.44$; Sections S9, S15).

Disattenuation. The observed convergent correlations understate true-score convergence because both instruments are measured with error. Using the classical correction $r_c = r / \sqrt{\text{rel}_{\text{ID}} \cdot \text{rel}_{\text{EQ-D}}}$ with the aggregate ID reliabilities observed here ($\text{ICC}(1, k) = .66$ per person-wave; split-half Spearman–Brown $r = .73$; $\text{ICC}(1, k) = .75$ in the neuroimaging cohort) and published EQ-D internal consistencies ($\alpha = .83\text{--}.90$ [Fresco et al., 2007]), the EQ-D correlation of $r = .36$ disattenuates to approximately .44–.49 (e.g., $.36 / \sqrt{.75 \times .90} = .44$; $.36 / \sqrt{.66 \times .83} = .49$). The MPoDT composite ($r = .41$, $\text{ICC}(1, k) = .65$) disattenuates similarly to roughly .5. These corrected values are descriptive context, not test statistics; all inferential claims in the paper use the observed correlations.

S5 Nomological: WASSA empathic concern (full)

Theory: decentering should support empathic concern (the compassionate, non-fused component) over and above personal distress. Table S6 reports the raw-essay view (robustness) and the pre-specified eligibility-filtered views; the both-filters view is primary for this mixed-register corpus. ID is itself positively related to personal distress ($r_{\text{dist}} = .28\text{--}.32$ across views), as expected given that empathic concern and personal distress themselves correlate ($r = .45$); the result is therefore incremental—ID predicts empathic concern net of distress (residual-on-distress $r = .30$; $\beta = .29$ controlling distress and length)—rather than a clean dissociation.

Table S6: WASSA empathic concern by view: zero-order empathy correlation (r_{emp}), the ID–personal-distress correlation (r_{dist}), residual-on-distress r , and β (controlling distress + log length).

View	r_{emp}	r_{dist}	resid distress	β	z	n
raw essay (robustness)	+.349	+.280	+.260	+.240	11.6	1860
mechanical	+.356	+.285	+.265	+.245	11.7	1831
LLM v2.2	+.397	+.317	+.304	+.286	13.1	1721
both (primary)	+.397	+.319	+.303	+.285	13.0	1705

S6 Nomological: Grossmann wise reasoning (full)

424 conflict-diary events (298 participants); standardized β from models with subject-clustered SEs, controlling log event length and survey wave. The primary subscale is significant; the three adjacent subscales the theory does not implicate are null, with 95% confidence intervals excluding standardized effects beyond $|\beta| = .16$ (script `20_nomological.py`).

Table S7: Grossmann wise-reasoning subscales ($n = 424$ events).

Subscale	β	95% CI	p
Search-for-Resolution (primary)	+.176	[+.048, +.304]	.007
Perspective-Taking (secondary)	+.101	[−.010, +.212]	.074
Compromise (null)	+.047	[−.066, +.160]	.415
Recognition of Limits (null)	+.019	[−.064, +.102]	.651
Expectation of Change (null)	−.009	[−.130, +.113]	.891

Eligibility filtering is a no-op here: the conflict-diary events are long first-person narratives, 100% scoreable under both the mechanical and v2.2 filters, so the Search-for-Resolution $\beta = .176$

is unchanged across all/mech/v2.2 views (ID-length $r = .38$, already controlled).

S7 Behavioral/incremental: MTurk (full)

ID $\times$ Big Five ($n = 418$): Agreeableness $+.41$, Openness $+.29$, Extraversion $+.14$, Conscientiousness $+.10$, Neuroticism $-.10$. Value-behavior coherence (controlling log word count and English fluency): overall partial $r = +.33$ (raw $+.37$), normative $+.27$, distinctive $+.20$ ($n = 417-418$). Incremental beyond the Big Five on VBC: base $R^2 = .263$, full $R^2 = .290$, $\Delta R^2 = +.027$, $F(1, 408) = 15.5$, $p = .0001$, $\beta_{ID} = +.18$. Narrative coherence: raw $r = +.33$, attenuating to partial $r = +.12$ ($p = .011$) controlling narrative log word count (treated as directional only).

Filter-view robustness (VBC). Value-behavior coherence is stable across eligibility views: partial r (controlling word count + English) = $.33$ (all) / $.32$ (mechanical) / $.34$ (v2.2) / $.32$ (mech $\wedge$ v2.2), all $p < .001$ ($n = 417$). Per-sentence retention was 74% (mechanical), 74% (v2.2), 57% (both). As the filters remove unscorable sentences the ID-length coupling falls from $r = .53$ (all) to $.39$ (both), yet the VBC partial is unchanged—so for this uniform first-person corpus the mechanical filter is sufficient and v2.2 adds little.

S8 Discriminant boundary: CMV (full)

For each ChangeMyView post, one comment earned a view-change award and one did not (4,311 pairs). Paired AUC for ID = $.55$, but text length alone gives AUC = $.62$: the apparent signal is length-confounded. After matching paired arguments to within 10% on length ($n = 462$ pairs), ID showed no positive discrimination (AUC = $.45$, 95% CI $[.40, .49]$); the small dip below $.50$ (sign test $p = .025$) is consistent with decentered, perspective-taking language being slightly less rhetorically forceful in debate, not evidence of systematic reverse prediction. Once length is controlled, ID carries no positive decentering signal in this off-construct register—the boundary the construct predicts.

S9 Neuroimaging cohort: known-groups, facets, and alternative accounts (full)

Table S8: Known-groups Hedges g vs. controls, and within-meditator paired comparison (standardized to controls).

Measure	meditators g	yoga g	within-med. ID vs. facet ($t(13)$, p)
Integrated Decentering	+1.83	+1.38	—
FFMQ Non-react	+1.56	+1.15	1.69, .116 (tie)
FFMQ Observe	+1.56	+1.51	2.60, .022
FFMQ Non-judge	+0.28	+0.55	5.41, $< .001$
FFMQ Describe	+0.25	+0.55	6.75, $< .001$
FFMQ Act-aware	+0.10	+0.23	7.52, $< .001$

Criterion regressions (ID + group + age + log word count, HC3): PSS $\beta = -.34$ ($p = .076$), SCS $\beta = +.34$ ($p = .021$), QOLI $\beta = +.23$ ($p = .28$), PCC-TPJ $\beta = +.34$ ($p = .27$).

Table S9: ID compared with human-coded TWS by outcome ($r_{\text{ID,TWS}} = .79$; per-outcome $n = 42\text{--}45$ after listwise deletion). TWS is an independently scored, cumulative expert score of ego development for the same sentence-completion protocols, with which ID shares a maturity-relevant axis (and same-protocol text). Williams’ t for two dependent correlations sharing a variable ($df = n - 3$) tests whether ID and TWS differ in how closely they track each outcome; p is two-tailed. The three psychological outcomes (PSS, QOLI, SCS) form the a-priori family; under Benjamini–Hochberg correction within that family SCS and QOLI hold at $q = .028$ (treating all six tests as one family gives $q = .056$). The three neural rows are exploratory and uncorrected. ΔR^2 partitions unique contributions in joint regressions adjusting for practice group. A model-derived maturity-score robustness check gives the same pattern (SCS $t = +2.71$, QOLI $+2.71$, PSS -1.12 ; Section S14). The PCC–TPJ edge is available for 43 of 45 participants (TPJ ROI coverage missing for two); PCC–IFC and PCC–dmPFC are complete ($n = 45$).

Outcome	n	r_{ID}	r_{TWS}	Williams’ t	p	$\Delta R^2_{\text{ID} \text{TWS}}$	$\Delta R^2_{\text{TWS} \text{ID}}$
PSS	43	−.44	−.33	−1.20	.236	.069	.013
QOLI	43	+.12	−.12	+2.59	.013	.085	.098
SCS	42	+.49	+.27	+2.46	.019	.112	.072
PCC–TPJ	43	+.24	+.17	+0.72	.473	.051	.000
PCC–IFC	45	+.54	+.52	+0.24	.809	.042	.014
PCC–dmPFC	45	+.20	+.24	−0.47	.641	.000	.034

Less stage-bound than the expert TWS score. ID is developmentally graded without reducing to discrete stage. A clarification first: TWS is a *cumulative* maturity score summed across responses, whereas developmental stage is assigned by applying threshold (cut-off) rules to the protocol; because the same total can arise from different response profiles and adjacent stages span overlapping score ranges, the score-to-stage map is many-to-one. A high correlation with TWS is therefore uninformative about whether ID recovers discrete stage; even $\text{ID} = \text{TWS}$ would not make $\text{ID} = \text{stage}$. In the neuroimaging cohort, which spanned conventional through postconventional developmental levels (stages 3–5, $\sim 16\%$ postconventional), ID correlated with human-coded developmental stage at $r = .69$ (Spearman $\rho = .64$, $n = 45$)—well below the $r = .95$ for TWS. Squaring these, discrete stage accounts for about 48% of ID’s variance against roughly 90% of TWS’s—roughly half—so the within-stage contrastive training attenuated, rather than eliminated, the stage association. ID nonetheless tracked TWS within stage (partial $r(\text{ID}, \text{TWS} \mid \text{stage}) = .63$): it indexes continuous reflective distance more than the stage label, while retaining the shared-axis correlation theory expects.

Decentering-relevant self-report net of developmental maturity. Within the neuroimaging cohort ($n = 43$ with complete self-report), ID predicted three decentering-relevant self-reports, and these associations were largely retained after controlling expert developmental maturity. ID correlated with the FFMQ Observe facet (a meta-awareness index) at $r = +.50$; the partial controlling TWS was $+.49$ ($p = .001$) and controlling discrete stage $+.48$ ($p = .001$), whereas TWS itself correlated with Observe only at $+.27$. For the FFMQ Non-reactivity facet (reduced reactivity), $r = +.50$; partial controlling TWS $+.41$ ($p = .006$) and controlling stage $+.52$ ($p < .001$), versus $+.33$ for TWS. For self-compassion (SCS), $r = +.49$; partial controlling TWS $+.47$ ($p = .002$) and controlling stage $+.54$ ($p < .001$), versus $+.27$ for TWS. Observe and Non-reactivity map onto two of the three decentering components in the Bernstein–Fresco model (meta-awareness and reduced reactivity) [Bernstein et al., 2015]. ID’s convergence with decentering-relevant self-report is thus

largely retained when developmental maturity is partialled out (Observe and self-compassion essentially unchanged; Non-reactivity attenuated modestly, .50 to .41), while maturity itself tracks these facets only weakly. Two caveats temper this: the indices are mindfulness *facets* rather than full decentering scales (e.g., the EQ-D [Fresco et al., 2007] or MPoDT), which were not co-measured with TWS in any cohort, and the estimates come from a single small cohort ($n = 43$). The pattern is therefore suggestive rather than decisive.

Beyond verbal IQ. ID correlated modestly with premorbid verbal IQ (AMNART-EVIQ; $r = .33$, $n = 33$), as expected for a language measure. Adding AMNART-EVIQ to each core criterion model (outcome $\sim$ ID + group + age + log word count) shifted the ID coefficient negligibly— $\Delta\beta_{ID} = -.010$ (PSS), $+.010$ (QOLI), $-.015$ (SCS), $-.000$ (PCC-TPJ), $-.001$ (PCC-IFC)—and AMNART itself predicted none of the outcomes (all $p > .29$). Consistent with the length-matched CMV boundary (Section S8) and the incremental-beyond-Big-Five result (Section S7), the criterion associations are not attributable to verbal fluency.

S10 Parent aging trial and secondary-analysis reporting

Registration and design. The parent study, “Neural and Cognitive Changes Associated With Mental Training in Older Adults,” was a single-site, two-arm, parallel randomized trial at Massachusetts General Hospital, Boston. It was registered at ClinicalTrials.gov on 11 December 2015 (NCT02628548) and ran from 18 January 2016 to 31 December 2021. The registry records 175 participants enrolled in the overall study. The primary trial report [Sevinc et al., 2021] records 1,472 screened and 146 randomized 1:1 to mindfulness training ($n = 72$) or cognitive-fitness training ($n = 74$). The discrepancy is one of study stage: registry enrollment, randomization in the primary report, and availability in this secondary analysis are distinct denominators. Full parent-trial methods are publicly available in the published report, with its published CONSORT flow diagram provided as supplementary material.

Participants, interventions, randomization, and masking. The parent trial recruited cognitively normal, meditation-naïve, right-handed English-speaking adults aged 65–80 through advertising, direct mail, and email lists; full medical, neuropsychiatric, MRI, cognitive, medication, and prior mind–body-practice criteria are reported by Sevinc et al. [Sevinc et al., 2021]. After baseline testing, the REDCap randomization module assigned participants 1:1 in permuted blocks of six stratified by gender; the allocation sequence was generated and concealed within this centralized module, so the personnel who enrolled participants and assigned them to the interventions could not view the sequence or upcoming allocations—each assignment was revealed by the module only at the point of randomization, after enrollment. The registry classifies masking as triple (participants, investigators, and outcome assessors); the primary report specifies that testing staff were blinded to assignment and that participants were told both programs promoted cognition to reduce expectation differences. The mindfulness program comprised eight weekly one-hour classes (45 minutes of practice plus 15 minutes of check-in/instruction) and 45 minutes of daily home practice; the active cognitive-fitness comparator matched class and home-practice time using puzzles. No concomitant care was delivered in either arm: beyond the assigned program, the trial provided no additional treatment or care. Mean attendance was 7.05 versus 6.64 of eight classes and mean home practice 4.01 versus 5.99 hours/week, respectively. Instructor qualifications and complete intervention content are reported in the primary report [Sevinc et al., 2021].

Outcomes and analysis status. The registered parent-trial primary outcomes were change in brain structure and brain activity at eight weeks; registered secondary outcomes included biomarkers and cognition. ID was not a registered or prespecified outcome. Every ID analysis in this paper is a post-hoc secondary analysis and was not preregistered. No sample-size calculation, interim analysis, stopping rule, or harms analysis was conducted for ID. ID analyses used all available observations meeting the requirements of each stated analysis, without imputation; exact statistical methods are described in Section 3.2. The parent trial completed follow-up as planned, and this report makes no claim about intervention benefits or harms.

Secondary-analysis flow. The published parent-trial CONSORT diagram reports that 67 of 72 mindfulness participants and 71 of 74 cognitive-fitness participants began treatment; five participants per arm discontinued treatment, one per arm was lost to follow-up, and 61 and 65, respectively, were analyzed for the parent trial’s cognitive outcome, with reasons reported in the diagram [Sevinc et al., 2021]. Sentence-completion text was available for 143 participants at one or more analyzed waves: 142 at baseline, 119 at 12 months, and 78 at 24 months. In the available secondary export, 120 participants were linked to one of the two randomized arms (58 mindfulness, 62 cognitive fitness); 23 records without an eligible two-arm link were excluded from arm comparisons. Complete baseline-to-12-month change was available for 53 mindfulness and 56 control participants; complete baseline-to-24-month change for 38 and 32, respectively. Reasons for missing long-term sentence-completion or practice data were not available in the secondary export. The 12-month arm contrast was $g = 0.02$ (95% CI $[-0.35, 0.40]$; mindfulness change $M = -0.10$, $SD = 0.68$; control $M = -0.12$, $SD = 0.71$); at 24 months it was $g = -0.15$ (95% CI $[-0.63, 0.32]$; mindfulness $M = -0.07$, $SD = 0.69$; control $M = 0.04$, $SD = 0.68$). Published parent-trial baseline characteristics and its original outcomes are reported in Tables 1–2 of Sevinc et al. [Sevinc et al., 2021]; the present report does not re-report those prespecified outcomes.

Table S10: Parent-trial and present secondary-analysis participant accounting. Dashes indicate that the public aggregate was not reported by arm.

Stage	Total	Mindfulness	Cognitive fitness
Registry enrollment	175	–	–
Randomized in the primary report	146	72	74
Begun treatment in the primary report	138	67	71
Parent cognitive outcome analyzed	126	61	65
Linked in the present secondary export	120	58	62
Complete baseline-to-12-month change	109	53	56
Complete baseline-to-24-month change	70	38	32

Table S11: Published parent-trial baseline characteristics (available cases), reproduced numerically from Table 1 of Sevinc et al. [Sevinc et al., 2021]. These describe the parent trial, not the present secondary analytic sample.

Characteristic	Mindfulness	Cognitive fitness
Age, years	70.2 ± 4.1 ($n = 70$)	71.0 ± 4.3 ($n = 75$)
Female	55.7% ($n = 70$)	53.3% ($n = 75$)
Education, years	16.7 ± 1.8 ($n = 69$)	16.7 ± 1.9 ($n = 74$)
Education, ISCED level	6.5 ± 1.0 ($n = 69$)	6.5 ± 1.1 ($n = 74$)

S11 Aging cohort: practice sensitivity and stability (full)

Practice follow-up was available for a subset, and exposure overlapped across arms: among reached participants, the mindfulness arm’s sitting-practice composite was 1.31 versus 0.35 in controls, while 12 of 20 reached controls reported some practice (Fisher $p = .08$ for any practice by arm). The available-case arm contrast was null at 12 months ($g = 0.02$, $p = .90$) and 24 months ($g = -0.15$, $p = .53$); outcome availability did not differ significantly by arm at 24 months (66% vs. 52%, $p = .14$). In a post-hoc joint model among 36 reached participants with 24-month change, reported practice was associated with change ($b = +0.70$, $p = .011$) whereas assigned arm was not ($p = .50$). This non-randomized exposure association is subject to selection and confounding. Exposure-based analyses and sensitivity checks are reported in Table S12; test–retest in the active-control arm was $r = .69$ ($n = 56$, 12 months), comparable at 24 months.

Table S12: Meditation dose–response and categorical contrasts (reached participants), new scorer.

Test	12 months	24 months
Dose, baseline-adj. partial r	+.26 ($p = .08$)	+.13 ($p = .40$)
Dose, Spearman ρ	+.29 ($p = .042$)	+.36 ($p = .017$)
Any practice vs. none, g	+.29 ($p = .35$)	+.80 ($p = .038$)
n (reached)	48	44

Baseline nomological net (discriminant dissociation). At baseline, ID (scoreable view, grand-average) tracks adaptive self-examination and wellbeing but not repetitive negative thought—the key discriminant test for a decentering measure (Table S13). Reflection, insight, and positive affect correlate positively; RRQ rumination and negative affect are null. (The EQ-D rumination subscale is weakly positive, $r = .23$, distinct from the RRQ repetitive-negative-thought scale.) This adaptive-versus-ruminative split is not what a relabeling of verbosity, distress, or general maturity would yield.

Table S13: Aging cohort baseline: ID vs. adaptive self-examination and repetitive negative thought. RRQ, Rumination–Reflection Questionnaire; PANAS, Positive and Negative Affect Schedule.

Criterion	r (95% CI)	n
RRQ self-reflection (adaptive)	+.42 [.27, .55]	133
SD-WISE insight	+.47 [.32, .60]	125
PANAS positive affect	+.39 [.23, .52]	133
FFMQ total mindfulness	+.31 [.15, .46]	132
PANAS negative affect	−.15 (<i>ns</i>)	133
RRQ rumination (RNT)	−.12 (<i>ns</i>)	133

S12 Reduced reactivity in coping responses (ECSC/COPING)

As an exploratory check on the reduced-reactivity facet of decentering, we scored 1,200 role-play responses from the Emotion and Coping Strategy Corpus (ECSC/COPING) [Troiano et al., 2024], in which a writer replies to a provocative statement under one of four assigned coping strategies—attack, reject, distance, contact—across five charged topics (reproducible scripts `08_ecsc_prep.py`—

10_ecsc_analyze.py). ID followed the predicted reactivity gradient, lowest for the reactive strategies and highest for engaged contact (attack 2.0 < reject 2.3 < distance 2.7 < contact 4.3). The reactive (attack/reject) versus non-reactive (distance/contact) contrast was substantial ($g = -0.47$) and held controlling response length and topic (partial $r = -0.24$, $p < .001$); a length-only classifier did not separate the strategies in the predicted direction (AUC 0.34, at or below chance), while adding ID raised it to 0.56, and a within-reactive placebo (attack vs. reject) was null—so the signal is the reactive stance, not verbosity. Within the corpus’s per-response emotion ratings, ID was selectively negative for reactive, aversive emotions (anger, disgust, contempt, dislike; partial $r \approx -.18$, net of length) yet positive for sadness (+.11) and null for guilt, shame, and regret, tracking a non-reactive stance rather than negative affect in general; the contrast persisted even when rated hostile-emotion content was partialled out (partial $r = -.17$), a conservative check since reduced reactivity is itself part of decentering. These role-play data are exploratory and correlational, but the pattern is the one a reduced-reactivity component of decentering predicts.

S13 Neuroimaging: connectivity detail and developmental gating (full)

As an exploratory extension we asked whether ID—derived purely from text—recovers a brain signature previously reported for this cohort. Singleton et al. [Singleton et al., 2021] analyzed the same 45 participants with an expert-coded developmental rating and the same PCC-seed resting-state connectivity, reporting (i) PCC–inferior-frontal-cortex (IFC) coupling as the primary whole-sample correlate and (ii) PCC–temporoparietal-junction (TPJ) and PCC–dorsomedial-prefrontal (dmPFC) coupling *sign-inflecting* at a developmental boundary (the transition to self-authoring), positive above and negative below. Because that boundary and those regions are fixed by prior work on this exact cohort, they provide a principled, hypothesis-constrained test not chosen here. However, because we reanalyze the same data and ID is highly correlated with that study’s developmental rating, it is a non-independent consistency check, not independent confirmation.

ID showed the same pattern from text (Figure S1). PCC–IFC, the whole-sample edge, correlated with ID at $r = .54$ ($\beta = .49$, $p = .009$, controlling group), robust across all eligibility-filter views ($\beta = .44$ –.50, all $p < .012$). PCC–TPJ was *developmentally gated*: at or below the boundary the ID–TPJ association was slightly negative ($r = -.22$), above it clearly positive ($r = +.50$), with an ID×stage interaction at $p = .046$ (uncorrected across the three edges, not surviving multiple-comparison correction; leave-one-out stable in sign), and the full-sample association correspondingly diluted ($\beta = .40$, $p = .14$)—the same pre/post averaging Singleton’s results imply. PCC–dmPFC showed the same gated pattern (interaction $p = .010$). The expert developmental rating reproduced the identical gating on the same edges (TPJ and dmPFC interactions both significant; IFC whole-sample), confirming that our connectivity measures are the regions Singleton reported and that the gating is a general developmental phenomenon shared by ID and the expert score.

We frame this as exploratory convergent evidence, not as an incremental claim. The gating is developmental and shared with the expert rating, so it is not decentering-unique; the cohort is small; developmental stage and practice are confounded here as in the original report; the TPJ variance partition into ID-specific versus shared components is unstable under leave-one-out, so we claim no ID-specificity beyond maturity at this edge; and connectivity was computed from the original 5-minute 1.5 T acquisition with the preprocessing of [Singleton et al., 2021], inherited without an independent motion or physiological-noise reanalysis.

The expert TWS rating shows the same gating on the same edges, confirming that the connectivity measures are the regions reported by Singleton et al. [Singleton et al., 2021] and that the

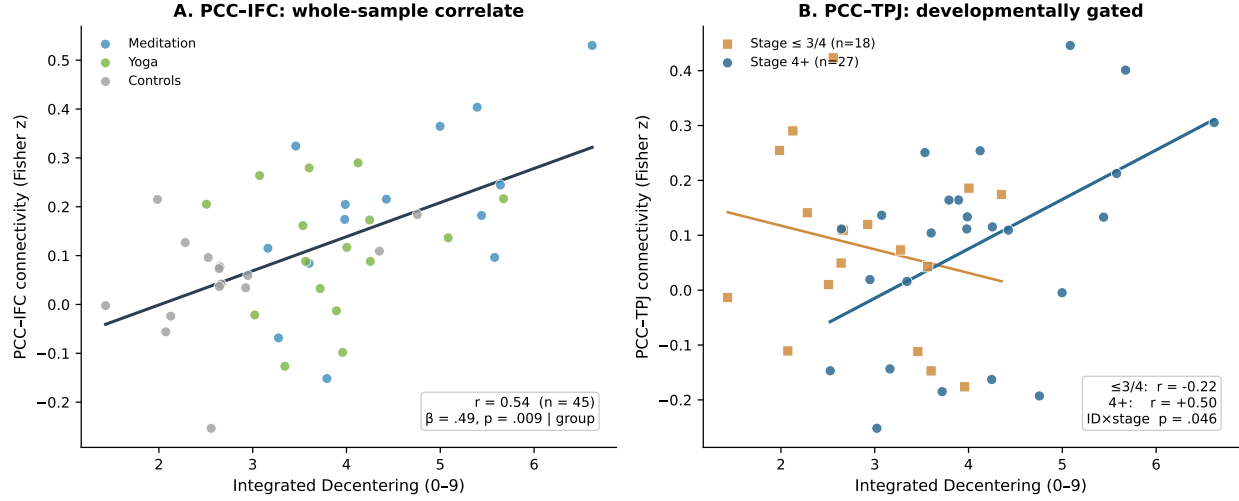


Figure S1: **Exploratory: ID recovers a developmentally gated PCC-connectivity pattern (neuroimaging cohort).** (A) PCC–IFC connectivity vs. ID, whole sample ($r = .54$; $\beta = .49$, $p = .009$ controlling group), coloured by practice group. (B) PCC–TPJ connectivity vs. ID, split at the human-coded developmental boundary (Stage $\leq 3/4$ vs. $4+$): the association is negative below and positive above (ID $\times$ stage $p = .046$), echoing Singleton et al. [Singleton et al., 2021]. Exploratory; small subgroups; see text for limits.

Table S14: PCC-seed connectivity by eligibility-filter view (group-controlled β , HC3).

Edge	all	mech	v2.2	both
PCC–IFC	.50**	.50**	.47**	.44*
PCC–TPJ	.40	.35	.41	.35
PCC–dmPFC	.20	.17	.25	.20

* $p < .05$, ** $p < .01$. PCC–IFC robust across views; PCC–TPJ/dmPFC not significant whole-sample (see gating below).

inflection is a general developmental phenomenon (shared by ID and TWS), not decentering-unique. A group-based contrast (control vs. practitioner) gives the same picture, as expected since stage and practice are confounded in this cohort (controls $\approx \leq 3/4$; practitioners ≥ 4). Within the post-threshold subgroup, the partition of PCC–TPJ variance into ID-specific versus shared components is unstable under leave-one-out, so we do not claim ID-specificity beyond maturity at this edge. Range restriction is present (post-threshold ID $M = 4.2$, $SD = 1.1$; pre $M = 2.9$, $SD = 0.8$).

S14 Robustness checks

Model-derived maturity (robustness for Section S9). Repeating the ID–TWS comparison with a model-derived maturity score in place of the human rating gave the same pattern: self-compassion and the quality-of-life sign divergence were significant (Williams’ $t = +2.71$ for both), while perceived stress was directionally consistent but not significant ($t = -1.12$). The human-coded rating is method-independent of the language model and is therefore reported in the main text.

Table S15: Developmental-stage moderation (human stage; pre $\leq 3/4$, $n = 18$; post 4+, $n = 27$). Subgroup r and ID $\times$ stage interaction p , for ID and for the expert TWS rating.

Edge	ID			TWS (expert)		
	r_{pre}	r_{post}	int. p	r_{pre}	r_{post}	int. p
PCC-IFC	+18	+62	.24	+02	+64	.31
PCC-TPJ	-.22	+50	.046	-.26	+39	.025
PCC-dmPFC	-.35	+41	.010	-.47	+45	.003

Filter-view robustness. Known-groups separation, the criterion regressions, and the ID–TWS correlation are stable across eligibility-filter views (e.g., meditator $g = 1.54\text{--}1.83$; $r_{\text{ID,TWS}} = .79\text{--}.81$ across views), confirming that none of the conclusions depends on a particular filtering choice.

Verbal IQ. The AMNART-adjusted criterion reanalysis is reported in Section S9: adding pre-morbid verbal IQ leaves the ID coefficients essentially unchanged and AMNART predicts none of the criteria, so verbal fluency does not carry the ID–criterion signal.

S15 Baseline comparison

ID is a *fixed* scorer with no parameters fit on these labels. We compare its simple correlation with each criterion against (i) a length-only baseline (log word count) and (ii) a TF-IDF + Ridge model trained directly on the criterion and evaluated out-of-sample (5-fold cross-validated predictions; GroupKFold by writer for the repeated-measures Grossmann corpus). The supervised baseline has every advantage—it is fit on the exact target in the exact corpus—yet ID matches it where it generalizes (WASSA), exceeds it where it does not (Grossmann, across writers, where the supervised model drops below the length baseline), and ID exceeds length on both nomological criteria, the supervised model only on WASSA (Table S18). The *same* ID scorer transfers across all three corpora, whereas each TF-IDF model is corpus- and target-specific. The one place a length baseline matches ID is the small MPoDT convergent sample (all stems); this reflects unscorable completions that carry only length—on the scoreable subset ID exceeds length and predicts MPoDT beyond it (Section S4). In Table S18, r_{ID} is the zero-order correlation; the length- and survey-wave-controlled Grossmann estimate is $\beta = .18$ (Section S6).

Surface linguistic-distancing baseline. Automated work on the neighbouring construct of psychological self-distancing indexes it through surface lexical markers—low first-person-singular pronoun use, more articles, more analytic/abstract language [Nook et al., 2017, Kross and Ayduk, 2017]. We tested whether ID merely recovers such markers by scoring every criterion text with the validated LIWC-22 dictionary [Boyd et al., 2022] (reproducible scripts `26a.make_liwc.csv.py` and `26b.liwc_distancing.py`) and forming (i) a Nook-style distancing composite (equal-weight z -scores: $-$ first-person-singular, $+$ articles, $+$ past- vs. $-$ present-focus, $+$ big words, $+$ analytic thinking, $+$ non-first-person reference, $+$ cognitive processing), (ii) LIWC’s analytic-thinking summary index alone [Pennebaker et al., 2014], and (iii) the best out-of-sample linear combination of the full distancing-relevant category set (5-fold cross-validated Ridge; GroupKFold by writer for Grossmann). In the two well-powered corpora the distancing index is essentially unrelated to the criterion (WASSA empathic concern: composite $r = .03$, Analytic $.07$, supervised-CV $.16$, vs. ID $.35$; Grossmann search-for-resolution: composite $-.07$, Analytic $-.03$, supervised-CV $-.02$, vs. ID $.20$), so

ID’s incremental contribution is large while the distancing index adds almost nothing (WASSA $\Delta R^2_{\text{ID}|\text{dist}} = .13$ vs. $\Delta R^2_{\text{dist}|\text{ID}} = .01$; Grossmann .04 vs. .00, ID $p = .001$ cluster-robust by writer; Table S16). On the small MPoDT convergent sample ($n = 35$) the distancing index does track the criterion ($r = .38$) and correlates with ID ($r = .55$)—unsurprising where the criterion is itself a self-report decentering scale—but $n = 35$ cannot partition unique variance (the supervised combination is not estimable at this size), so the discriminant evidence rests on the two large corpora. ID is not a relabeling of surface distancing markers.

Table S16: Surface linguistic-distancing baseline (LIWC-22) vs. ID. r_{dist} is the equal-weight Nook-style distancing composite; r_{Analytic} is LIWC’s analytic-thinking summary index alone [Pennebaker et al., 2014]. ΔR^2 partitions unique variance in the joint model (ID + composite). In the two large corpora the distancing index is near zero with the criterion and adds almost nothing beyond ID; at MPoDT ($n = 35$) the index tracks the criterion but the sample is too small to partition unique variance (increments not significant).

Corpus / criterion	n	r_{ID}	r_{dist}	r_{Analytic}	$\Delta R^2_{\text{ID} \text{dist}}$	$\Delta R^2_{\text{dist} \text{ID}}$
WASSA: empathic concern	1860	+.35	+.03	+.07	+.13	+.01
Grossmann: search-for-resolution	424	+.20	−.07	−.03	+.04	+.00
MPoDT composite	35	+.41	+.38	+.28	+.06	+.03

Zero-shot LLM baselines. A second baseline asks whether the contrastive fine-tuning adds anything over simply prompting a capable model. We zero-shot six models with a strong decentering rubric (temperature 0, reasoning disabled, per-completion scoring identical to ID’s pipeline; reproducible script `19_zeroshot.py`): the *same* two base models ID fine-tunes (Qwen3.5-27B + Qwen3.6-27B; their average is the untuned base ensemble—a direct ablation), two mid-tier models (GPT-4.1-mini, Claude-3.5-Haiku), and the current frontier models (GPT-5.5, Claude Opus 4.8). All comparisons are matched: every model is scored on the same stems and aggregated identically, on both the all-stems and scoreable (mech $\wedge$ v2.2) views. On the well-powered EQ-D ($n = 124$), ID substantially exceeds the zero-shot models on either view (ID .31 all / .33 scoreable): the untuned base ensemble (.07/.08), the mid-tier models ($\leq .14$), and the current frontier models, which do no better and approach zero (GPT-5.5 .08/.01, Claude Opus 4.8 .02/−.01). Across the zero-shot, few-shot, and reasoning-enabled prompting tested, capability and scale did not substitute for the targeted fine-tuning. On the small MPoDT sample ($n = 35$) estimates are noisy and not distinguishable (GPT-5.5 .48/.46 brackets ID’s .41/.48; Opus 4.8 is low at .21/.23), so the powered EQ-D is the reliable comparison. Beyond accuracy, zero-shot frontier scoring is non-reproducible (model versions drift) and uncalibrated, whereas ID is a small, fixed, deterministic, released instrument. All prompted models were queried through the OpenRouter API at temperature 0 with reasoning disabled (the reasoning-enabled check used reasoning effort “low”), one completion per call; the system prompt and scoring rubric are in the reproducibility scripts (`19_zeroshot.py`, `19d_fewshot.py`, `19c_zeroshot_or_reasoning.py`). Because prompted-model outputs drift with provider updates, we pin the model identifiers and query dates: the zero-shot EQ-D comparison used `qwen/qwen3.5-27b`, `qwen/qwen3.6-27b`, `openai/gpt-4.1-mini`, `anthropic/claude-3.5-haiku`, `anthropic/claude-opus-4.8`, and `openai/gpt-5.5` (queried 2026-05-29); the reasoning-enabled runs of the two most capable models were made on 2026-06-12 and the few-shot runs on 2026-07-04. Per-completion scored outputs for every model are retained in the reproducibility package.

Pretraining exposure. Several validation corpora (WASSA, CMV, the MTurk sample) are public and may occur in the pretraining data of the Qwen base models; the private MGH cohorts and the author-collected MPoDT sample cannot. Three features of the design bound this concern. First, the scoring task is judging a passage presented in isolation, not recalling it: no criterion label accompanies any text at inference, and the public releases pair labels with texts only as research-download metadata. Second, the untuned base ensemble—identical pretraining, no contrastive fine-tuning—shows no convergent signal ($r = .07$), so pretraining exposure alone demonstrably does not produce the validation pattern; the signal enters with fine-tuning, whose material is non-public and disjoint from every validation sample (Section S1). Third, the convergent (EQ-D), reliability, and practice-linked results all come from the private clinical cohorts. Memorization is therefore an implausible account of the pattern.

Few-shot frontier check. Because zero-shot prompting might understate what in-context examples can elicit, we re-ran the two frontier models on the identical EQ-D pipeline with the same rubric plus six author-written anchor exemplars spanning the 0–9 scale, presented as in-context dialogue turns (scoring-manual-style anchors; synthetic, so no training or validation text enters the prompt; temperature 0, reasoning disabled; reproducible script `19d_fewshot.py`). Few-shot prompting did not produce convergent signal: Claude Opus 4.8 $r = .06$ (zero-shot $.02$) and GPT-5.5 $r = .08$ (zero-shot $.08$), against ID’s $r = .31$ on the matched wave. We did not tune the exemplar set; as with the other prompted baselines, some untested prompt could conceivably do better, but rubric, exemplars, and reasoning each failed to move the frontier models off floor on this test.

Reasoning-enabled frontier check. Because the two frontier models are reasoning-first systems, we re-ran them with reasoning enabled (pinned OpenRouter API, reasoning effort low, one completion per call, 3,042 scoreable EQ-D completions across 123 persons; reproducible script `19c_zeroshot_or_reasoning.py`). On the identical scoreable view, reasoning did not produce convergent signal: GPT-5.5 $r \approx .00$ ($p = .99$) and Claude Opus 4.8 $r = .03$ ($p = .72$), against ID’s $r = .33$ on the same view (the reasoning-disabled estimates on this view are $.01$ and $-.01$). This comparison was run on the powered EQ-D test only; the small MPoDT sample ($n = 35$) cannot separate scorers at any operating point, as the reasoning-disabled estimates already show.

Table S17: Zero-shot LLM decentering scoring vs. ID, per-person, on the convergent criteria. Models span the open base (Qwen3.5/3.6-27B), mid-tier (GPT-4.1-mini, Claude-3.5-Haiku), and the current frontier models (GPT-5.5, Claude Opus 4.8). EQ-D column is the matched baseline (t1) wave, all-stems; the scoreable subset gives the same ranking (ID $.33$ vs. $\leq .12$ for every zero-shot model, frontier models ≈ 0). ID grand-average across waves is $.36$. Enabling reasoning, or few-shot prompting with six anchored exemplars, leaves the two frontier models near zero (Opus 4.8 $.06$, GPT-5.5 $.08$ few-shot; see text).

Scorer	MPoDT (all)	MPoDT (scoreable)	EQ-D (t1)
Qwen3.5-27B (zero-shot)	.37	.30	.08
Qwen3.6-27B (zero-shot)	.46	.38	.05
base ensemble (untuned)	.42	.33	.07
GPT-4.1-mini (zero-shot)	.37	.33	.13
Claude-3.5-Haiku (zero-shot)	.41	.38	.14
GPT-5.5 (zero-shot, <i>frontier</i>)	.48	.46	.08
Claude Opus 4.8 (zero-shot, <i>frontier</i>)	.21	.23	.02
ID (LoRA-tuned ensemble)	.41	.48	.31

Table S18: ID (fixed) vs. length-only and supervised TF-IDF+Ridge (cross-validated) on text criteria.

Corpus / criterion	n	r_{ID}	r_{length}	$r_{\text{TF-IDF (CV)}}$
WASSA: empathic concern	1860	+.35	+.14	+.36
Grossmann: search-for-resolution	424	+.20	+.12	+.01
MPoDT composite	35	+.41	+.45	+.17

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PARENTTRIALCONSORTFLOWATTRIBUTED.pdf](#)
- [CONSORT2025COMPLETED.pdf](#)